University of Cape Town's WMT22 System: Multilingual Machine Translation for Southern African Languages

Khalid N. Elmadani Francois Meyer Jan Buys
Department of Computer Science
University of Cape Town
{ahmkha009,myrfra008}@myuct.ac.za, jbuys@cs.uct.ac.za

Abstract

The paper describes the University of Cape Town's submission to the constrained track of the WMT22 Shared Task: Large-Scale Machine Translation Evaluation for African Languages. Our system is a single multilingual translation model that translates between English and 8 South / South East African Languages, as well as between specific pairs of the African languages. We used several techniques suited for low-resource machine translation (MT), including overlap BPE, back-translation, synthetic training data generation, and adding more translation directions during training. Our results show the value of these techniques, especially for directions where very little or no bilingual training data is available.[1]

1 Introduction

Southern African languages are underrepresented in NLP research, in part because most of them are low-resource languages: It is not always possible to find high-quality datasets that are large enough to train effective deep learning models (Kreutzer et al., 2021). The WMT22 Shared Task on Large-Scale Machine Translation Evaluation for African Languages (Adelani et al., 2022) presented an opportunity to apply one of the most promising recent developments in NLP — multilingual neural machine translation — to Southern African languages. For many languages, the parallel corpora released for the shared task are the largest publicly available datasets yet. For some translation directions (e.g. between Southern African languages), no parallel corpora were previously available.

In this paper we present our submission to the shared task. Our system is a Transformer-based encoder-decoder (Vaswani et al., 2017) that translates between English and 8 South / South East African languages (Afrikaans, Northern Sotho,

Shona, Swati, Tswana, Xhosa, Xitsonga, Zulu) and in 8 additional directions (Xhosa to Zulu, Zulu to Shona, Shona to Afrikaans, Afrikaans to Swati, Swati to Tswana, Tswana to Xitsonga, Xitsonga to Northern Sotho, Northern Sotho to Xhosa). We trained a single model with shared encoder and decoder parameters and a shared subword vocabulary.

We applied several methods aimed at improving translation performance in a low-resource setting. We experimented with BPE (Sennrich et al., 2016b) and overlap BPE (Patil et al., 2022), the latter of which increases the representation of low-resource language tokens in the shared subword vocabulary. We used initial multilingual and bilingual models to generate back-translated sentences (Sennrich et al., 2016a) for subsequent training.

First, we trained a model to translate between English and the 8 Southern African languages. Then we added the 8 additional translation directions and continued training. For some of these additional directions no parallel corpora were available, so we generated synthetic training data with our existing model. By downsampling some of the parallel corpora to ensure a balanced dataset, we were able to train our model effectively in the new directions, while retaining performance in the old directions.

We describe the development of our model and report translation performance at each training stage. Our final results compare favourably to existing works with overlapping translation directions. While there is considerable disparity in performance across languages, our model nonetheless achieves results that indicate some degree of effective MT across all directions (most BLEU scores are above 10 and most chrF++ scores are above 40). We also discuss our findings regarding techniques for low-resource MT. We found overlap BPE and back-translation to improve performance for most translation directions. Furthermore, our results confirm the value of multilingual models, which proves critical for the lowest-resource languages.

---

[1]Our model is available at https://github.com/Khalid-Nabigh/UCT-s-WMT22-shared-task.

2 Background

2.1 Multilingual Neural Machine Translation (MNMT)

Multilingual models help low-resource languages (LRLs) by leveraging the massive amount of training data available in high-resource languages (HRLs) (Aharoni et al., 2019; Zhang et al., 2020). In the context of Neural Machine Translation, a multilingual model can translate between more than two languages. Current research in MNMT can be divided into two main areas: training language-specific parameters (Kim et al., 2019; Philip et al., 2020) and training a single massive model that shares all parameters among all languages (Fan et al., 2020; NLLB Team et al., 2022). Our work lies in the second category, as we are building a single multilingual translation system by exploring back-translation and different vocabulary generation approaches.

2.2 Back-Translation

Given parallel sentences in two languages A and B (A_b, B_a), with goal of training a model that translates sentences from A to B ($A \rightarrow B$). Back-translation works as follows: First, one trains a ($B \rightarrow A$) model using the available (A_b, B_a) data. Then the B_a sentences are passed to the model to regenerate A_b. This model's output (A_b') is then considered as additional synthetic parallel data (A_b', B_a). The final step of back-translation is training an ($A \rightarrow B$) translation model using (A_b', B_a) as parallel data. The motivation behind back-translation is that the noise added to the A_b' sentences from regeneration increases the model's robustness (Edunov et al., 2018). The same approach can be extended to multilingual models (Liao et al., 2021).

2.3 Overlap-based BPE (OBPE)

Byte Pair Encoding (BPE) is a vocabulary creation method that relies on n-gram frequency (Sennrich et al., 2016b). The starting point is a character-based vocabulary. At each step, the BPE algorithm identifies the two adjacent tokens with the highest frequency, joins them together as a single token, and adds the new token to the vocabulary. The dataset is then restructured based on the expanded vocabulary. In the case of multilingual training, a single BPE vocabulary can handle all languages by running the BPE algorithm on the union of the data from all the languages. However, when con-

Language Pairs	WMT22_african
eng-sna	8.7M
eng-xho	8.6M
eng-tsn	5.9M
eng-zul	3.8M
eng-nso	3M
eng-afr	1.6M
eng-tso	630K
eng-ssw	165K
xho-zul	1M
zul-sna	1.1M
sna-afr	1.6M*
afr-ssw	165K*
ssw-tsn	85K
tsn-tso	285K
tso-nso	212K
nso-xho	200K

Table 1: Number of available parallel sentences for all language pairs. * indicates that no data is available for these pairs and the number represents the amount of synthetic data we generated.

Language Family	L_{HRL}	L_{LRL}
Germanic	English(eng)	Afrikaans(afr)
Nguni	Xhosa(xho)	Zulu(zul), Swati(ssw)
Sotho-Tswana	Tswana(tsn)	Sepedi(nso)
Bantu	Shona(sna)	Xitsonga(tso)

Table 2: The languages included in our translation system, grouped by language family and whether they are used as L_{HRL} or L_{LRL} for the OBPE algorithm.

structing a multilingual vocabulary, BPE will prefer frequent word types, most of which are from HRLs, leaving a smaller proportion of the vocabulary for words from LRLs.

Overlap-based BPE (OBPE) is a modification to the BPE vocabulary creation algorithm which enhances overlap across related languages (Patil et al., 2022). OBPE takes into account the frequency of tokens as well as their existence among different languages. Given a list of HRLs (L_{HRL}) and LRLs (L_{LRL}), OBPE tries to balance cross-lingual sharing (tokens shared between HRLs and LRLs) and individual languages' representation. The optimal OBPE vocabulary for a set of languages from different families is produced by considering the highest resource language from each family as L_{HRL} and the rest of the languages as L_{LRL}.

3 Datasets

The WMT22 dataset is released along with the shared task. It contains bitext for 248 pairs of African languages, referred to as `WMT22_african`.[2] We use `WMT22_african` for both training and validation; the first 3 000 sentences from each language pair is reserved for validation and the rest for training. Table 1 shows available number of sentences for each language pair. No data was provided for Shona-Afrikaans and Afrikaans-Swati, so we generated synthetic data for these translation directions (see section 4.2.1). For testing, we used the Flores dev set, which contains 997 parallel sentences for each language pair. Additionally, we report the results of the final translation system as evaluated by the shared task organizers on a hidden test set.

3.1 OBPE

We trained BPE and OBPE tokenizers using the $eng \leftrightarrow$ LRL data only (the first 8 rows of table 1). The vocabulary size for both BPE and OBPE is set to 40K. For OBPE, the L_{HRL} contains the highest-resource language from each language family (eng, xho, tsn, sna), while L_{LRL} includes the rest of the languages (see table 2). We used Patil et al.'s (2022) implementation for both BPE and OBPE. This implementation is based on the Hugging Face Tokenizers library.[3]

4 Methodology

In this work, we only focus on South and South East African languages, their translation to/from English, and eight translation directions between these languages. We divided the training of the translation system into two stages. In the first stage, we trained a multilingual model for translating from all LRLs to English and vice versa. To incorporate the translation directions between LRLs into the system, we did further training on the translation model from stage 1. We divided the training process into stages instead of training the model in one session due to computational resource constraints. Both stages are explained in more detail below.

All models were trained with the `Fairseq` toolkit (Ott et al., 2019). We used the `transformer-base` architecture (Vaswani et al., 2017) for training all bilingual models. We base

Data	Δ
sna-eng	0.1
xho-eng	0.2
tsn-eng	−0.2
zul-eng	−0.7
nso-eng	0.3
afr-eng	0.0
tso-eng	0.0
ssw-eng	0.3
eng-sna	0.1
eng-xho	−0.2
eng-tsn	0.2
eng-zul	0.1
eng-nso	−0.2
eng-afr	0.0
eng-tso	−0.2
eng-ssw	0.0

Table 3: BLEU score differences between the OBPE multilingual model (13th epoch) and the BPE multilingual model (10th epoch) on Flores dev set. We stopped training the BPE model at this point as the OBPE model is computationally more efficient. The translation directions are sorted based on the available amount data.

the multilingual models on the BART architecture (Liu et al., 2020), using Tang et al.'s (2021) implementation and hyperparameters, including adding a token to indicate the source language before the input sentence and a token for the target language before the output sentence.

4.1 Stage 1: Translation Between LRLs and English

We used BPE and OBPE vocabularies to train two multilingual models for all directions between English and LRLs. Bilingual models were trained for each translation direction using a single vocabulary for each model. Finally, we performed back-translation for all directions using the model with the highest BLEU score in each case.

4.1.1 Multilingual Training

Multilingual models generally have more parameters and require more training time and computational resources than bilingual models. Computational constraints prevented us from fully training two multilingual models and then doing back-translation from them. Subsequently we used BPE and OBPE vocabularies to train two multilingual models till the 10th and 13th epochs, respectively. At this point, we found that the difference in trans-
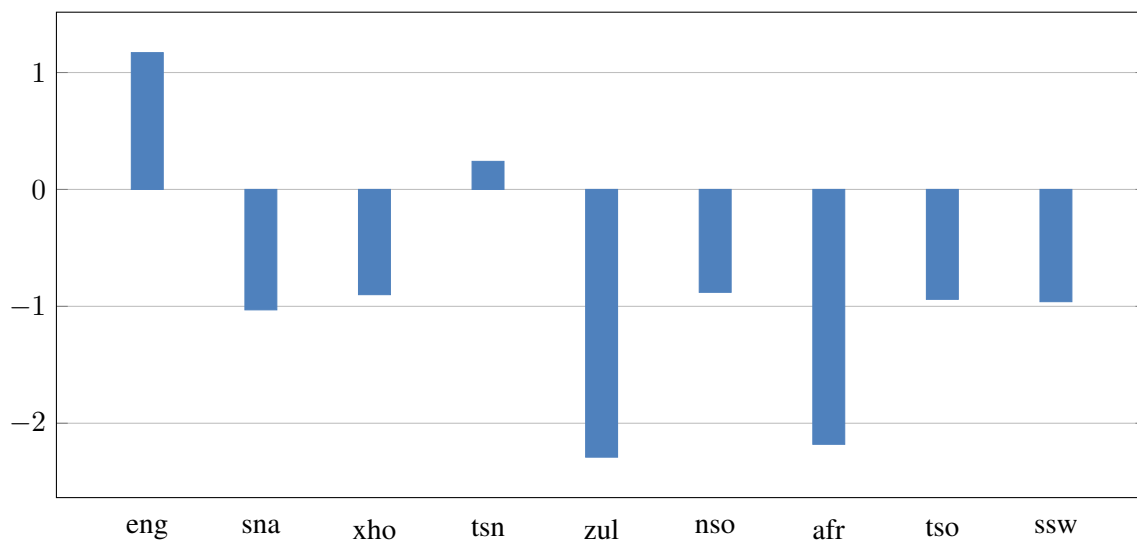
Figure 1: The change in the number of tokens in the training set per language when using OBPE instead of BPE. Less training tokens correspond to better a representation of a language in the shared subword vocabulary, so negative percentage changes reflect an improvement in low-resource language representation.
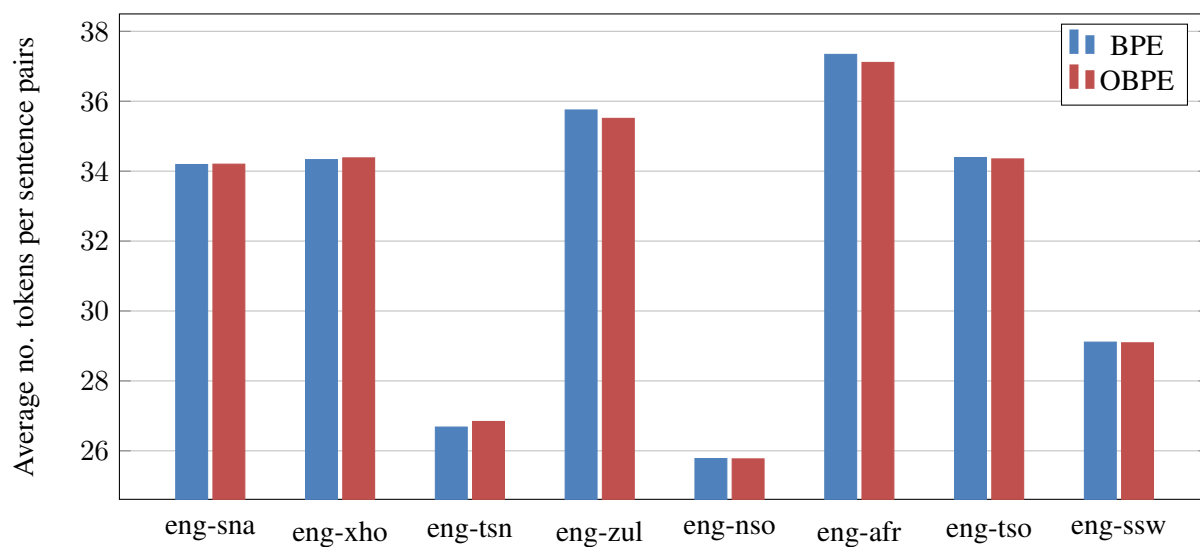

Figure 2: The average number of tokens per sentence pair for all language pairs with English, comparing BPE and OBPE vocabularies. More tokens lead to slower training.

lation quality between the two models is negligible (see table 3). However, the OBPE model is slightly faster in training and represent LRLs better. A language l is represented better in vocabulary V_1 than V_2 if V_1 contains more subword tokens from l than V_2. The total number of tokens in l's training data will influence its representation in the vocabulary. Reducing the number of tokens in the training sentences requires increasing the vocabulary capacity. Therefore, fewer tokens in the training data corresponds to a better vocabulary representation. We are interested in comparing BPE and OBPE's vocabulary representation for all languages. We used the following formula to measure the relative change in the number of training tokens when using OBPE instead of BPE,

$$\text{change}_l = \frac{T^l_{\text{OBPE}} - T^l_{\text{BPE}}}{T^l_{\text{BPE}}}\% \qquad (1)$$

where T^l_{BPE} and T^l_{OBPE} are the total number of tokens in language l's training data when using BPE and OBPE vocabularies, respectively. Figure 1 shows the change in number of training tokens for all languages. The negative sign in the figure indicates that OBPE represents the language better than BPE. It can be clearly seen that OBPE represents most LRLs better than BPE.

As we are training autoregressive models, the training speed depends on the number of target tokens, which is controlled by the target language representation in the subword vocabulary. Therefore we use the average number of tokens per training example for each language pair (*eng-l*) as a proxy for training speed. Fewer tokens leads to faster training. Both source and target tokens are included, as we are training the model to translate in both directions:

$$\text{AvgTokens}^V_{eng-l} = \frac{\text{Tok}^l_{eng-l} + \text{Tok}^{eng}_{eng-l}}{N_{eng-l}} \qquad (2)$$

where $\text{AvgTokens}^V_{eng-l}$ indicates the average number of tokens in one training example from the $eng - l$ dataset using V vocabulary. Tok^l_{eng-l} and Tok^{eng}_{eng-l} represent the total of l and eng tokens, respectively, in the $eng - l$ dataset, while N_{eng-l} represents the number of training examples in the same dataset. Figure 2 shows the average number of training tokens in each language pair when using BPE and OBPE vocabularies. We observe that training with OBPE is slightly faster than training

with BPE. The speed difference is higher for languages that are better represented by OBPE (see figure 1).

For these two reasons, and due to time and resources constrains, we chose to continue with training the OBPE multilingual model only.

4.1.2 Bilingual Training

Multilingual models often harm performance on high-resource languages compared to their bilingual counterparts (Yang et al., 2022). For back-translation, we used bilingual models for the subset of language pairs where this happens. We had two translation directions for each language (from/to English) and two vocabulary options (BPE/OBPE) for each direction. We ended up with 32 bilingual models.

All bilingual models were trained on either an `Nvidia` A100 full card (40GB) or a division of half a card (20GB) for 45 epochs with a batch size of 12 288 tokens. The training time depends on the language pairs, but the highest-resource language pair took three days of training.

4.1.3 Back-Translation

For each translation direction, we choose one of the following models for generating back-translation sentences: OBPE bilingual, BPE bilingual, and the 17th epoch checkpoint from the OBPE multilingual model. The selection is based on the models' performance on the Flores dev set, as measured by their BLEU score. We generated the back-translation sentences from the available parallel data only; no additional monolingual data was used. Results from table 4 show the performance of those three models. It can be seen that bilingual models are performing better in both directions of the higher-resource language pairs and for *eng-afr*. We discuss the results in more details in section 5.

We trained the OBPE multilingual model until the 17th epoch. That checkpoint was then used to generate back-translation data for the directions where the multilingual models outperform bilingual ones. Due to resources and time constraints, we started training the back-translation multilingual model from the 17th epoch checkpoint of the OBPE multilingual model. The OBPE multilingual model continued training regularly from the 17th epoch.

We ran all multilingual experiments on 2 `Nvidia` A100 cards (40GB each). One epoch of back-translation or OBPE multilingual models took 16

hours. Both models trained for 45 epochs with a batch size of 16 384 tokens, leading to a total training time of 30 days for each model.

After training both multilingual models, we had four models for each translation direction; two bilingual and two multilingual models.

4.2 Stage 2: Translation Between LRLs

At this stage we found that our models showed adequate performance in the English-centric directions (similar evaluation scores to existing works with overlapping translation directions). The goal of the next stage was to add new translation directions between specific LRLs. Our best multilingual model at this point (based on BLEU scores in the English-centric directions) was the OBPE-based model that was partially trained on back-translated data. Therefore we selected this model to continue training in the new directions. The model trained for an additional 39 epochs on a training set covering the old and new directions (details in section 4.2.2). This took 9 days on a full `Nvidia A100` card (40GB), at which point validation performance had stopped improving. This resulting model is the system we submitted to the shared task.

4.2.1 Synthetic training data

As shown in table 1, the translation directions between LRLs (new directions) generally had smaller datasets than the directions from/to English (old directions). In fact, two of the new directions (Shona to Afrikaans and Afrikaans to Swati) had no parallel corpora at all. To add these two directions to the model, we generated partially synthetic training data using the available English-centric parallel corpora. Using our multilingual model, we translated the English sentences in the English-Afrikaans corpus to Shona, and the English sentences in the English-Siswati corpus to Afrikaans. This produced parallel corpora for Shona-Afrikaans and Afrikaans-Siswati, where the target sentences were real and the source sentences were synthetic.

4.2.2 Balancing parallel corpora

The challenge in adding new translation directions is to strike a balance between gaining performance in the new directions, while ensuring that performance in the old directions does not deteriorate in the process. For this stage our model was trained on parallel corpora in the old and new directions. Including training data for the old directions ensures that the model does not lose its translation abilities

for these directions. However, the parallel corpora for the old directions are on average much larger than those of the new directions. Therefore training on such an unbalanced dataset would likely result in suboptimal performance for new directions.

To counter this, we downsampled the training data for the old directions to match the corresponding corpora in the new directions in order to balance the model's exposure to the old and new directions during training. For example, to balance Xhosa to Zulu training (1M sentences), we trained on 1M sentences only from both the English to Zulu and the Xhosa to English corpora. Therefore the encoder is trained for Xhosa balancing the Xhosa-English and Xhosa-Zulu data, while the decoder is trained for Zulu balancing the English-Zulu and Xhosa-Zulu setting.

Another potentially better approach is upsampling the training data for new directions. This technique would ensure that the model is exposed to all training data of old directions. However, we did not explore this due to time constraints.

5 Results

We primarily used BLEU score for evaluating all models on the Flores dev set. The final test set evaluation by the shared task organizers additionally used sentence piece BLEU (spBLEU) and chrf2.

5.1 Translation Between English and LRLs

Table 4 shows our results on the translation between English and LRLs. For each translation direction, we selected the best model among the two bilingual models and the 17th epoch checkpoint of the OBPE multilingual to perform back-translation. Although the multilingual model was trained only for 17 epochs, it outperformed the fully trained bilingual models in some language pairs. Most of these pairs are resource-poor ($eng \leftrightarrow nso, tso, ssw$). The exception of this finding was the translations between English and Afrikaans. These two languages are from the same family, so we hypothesize that the bilingual models did not need help from other resource-rich pairs or additional training examples to translate between the two languages. The training data of resource-richer language pairs ($eng \leftrightarrow xho, zul, tsn$) were sufficient to train good bilingual models.

After we fully trained both OBPE and OBPE+back-translation multilingual models, the OBPE model performed better than the

Data	Bi-BPE	Bi-OBPE	M-OBPE@17	M-OBPE	M-OBPE+back	M-OBPE-final
sna-eng	19.1	<u>19.6</u>	17.7	19.1	18.1	**19.5**
xho-eng	26.2	<u>26.9</u>	24.2	26.3	26.5	**27.5**
tsn-eng	11.8	11.9	<u>18.1</u>	19.2	16.1	**20.3**
zul-eng	<u>28.7</u>	28.2	26.4	28.6	**30.0**	**30.0**
nso-eng	12.9	14.6	<u>23.1</u>	25.5	22.9	**26.9**
afr-eng	47.5	**48.5**	41.8	45.0	46.4	44.8
tso-eng	1.1	3.3	<u>17.2</u>	18.8	16.9	**20.7**
ssw-eng	0.7	0.9	<u>19.4</u>	21.3	18.0	**23.0**
avg	18.5	19.2	23.5	25.5	24.4	**26.6**
eng-sna	<u>10.1</u>	9.9	9.3	10.0	10.1	**10.3**
eng-xho	12.3	<u>12.6</u>	10.9	11.8	**12.7**	12.1
eng-tsn	10.2	9.6	<u>16.5</u>	17.8	17.8	**18.2**
eng-zul	<u>14.9</u>	14.3	12.6	14.2	**15.1**	15.0
eng-nso	9.8	10.4	<u>20.3</u>	22.1	22.3	**23.1**
eng-afr	**37.2**	35.8	32.3	34.1	36.2	35.6
eng-tso	0.7	0.9	<u>12.8</u>	14.5	15.0	**16.9**
eng-ssw	0.7	0.9	<u>6.2</u>	6.9	7.0	**7.7**
avg	12	11.8	15.1	16.4	17	**17.4**

Table 4: BLEU scores on Flores dev set for translating between English and LRLs. The translation directions are sorted based on the available amount data. Bi-BPE and Bi-OBPE are the BPE and OBPE bilingual models, respectively. M-OBPE@17 is the 17th epoch checkpoints of the OBPE multilingual model, while M-OBPE is trained for 45 epochs. M-OBPE+back and M-OBPE-final are the OBPE with back-translation multilingual models before and after continued training for translation between LRL, respectively. <u>underline</u> indicates the model we used for back-translation. **Bold** represents the best overall model.

Data	M-OBPE+back	M-OBPE-final
xho-zul	1.5	**11.2**
zul-sna	1.9	**8.8**
sna-afr	1.9	**12.2**
afr-ssw	1.3	**4.9**
ssw-tsn	2.0	**14.5**
tsn-tso	2.1	**13.6**
tso-nso	2.4	**13.2**
nso-xho	1.7	**8.2**
avg	1.8	**10.8**

Table 5: BLEU scores on Flores dev set for translating between LRLs. M-OBPE+back and M-OBPE-final are the OBPE multilingual models with back-translation before and after continued training for translation between LRL, respectively. M-OBPE-final is the system we submitted for the shared task. **Bold** represents the best results.

back-translation model in most directions with English as a target language, namely, $sna, tsn, nso, tso, ssw \rightarrow eng$. However, for the three eng generation directions where the back-translation model performed similarly or better than the OBPE model ($xho, zul, afr \rightarrow eng$),

the back-translation data was generated from the bilingual models, not the OBPE multilingual model. This synthetic data contains actual English sentences and synthetic LRLs sentences. These translation pairs were relatively resource-rich. In contrast, most of the remaining pairs were resource-poor, and their back-translation data was generated from the partially trained OBPE multilingual model. These results show that although the 17th epoch checkpoint of the OBPE multilingual model was better than bilingual models in resource-poor language pairs, it was not yet good enough for generating text in LRLs. This led to a performance drop for the back-translation model on most of the eng generation directions compared to the OBPE multilingual model.

On the other hand, the back-translation model outperformed the OBPE model in all directions translating into LRLs. These directions require synthetic English sentences and actual LRLs sentences for back-translation. A plausible explanation for this is that learning to translate to English is easier than translating to LRLs for both bilingual and multilingual models.

Data	BLEU	spBLEU	CHRF2++	ΔCHRF2++
sna-eng	18.7	22.1	42.9	5.5
xho-eng	24.3	26.8	47.7	5.6
tsn-eng	19.8	22.1	42.6	7.7
zul-eng	26.7	28.5	49.3	6.5
nso-eng	26.5	28	48.1	9.4
afr-eng	44.7	46.4	66	9
tso-eng	20.3	21.8	41.9	8.8
ssw-eng	21.5	23.5	43.8	7.9
avg	25.31	27.4	47.79	7.55
eng-sna	10.3	17.6	41.1	2.9
eng-xho	9.4	18.6	42.5	3.4
eng-tsn	18.8	19.7	43	5
eng-zul	11.9	22.8	46.1	3.4
eng-nso	22.7	24.1	47.8	4
eng-afr	35.9	40.5	62.2	3.6
eng-tso	15.8	17.9	41.5	4.8
eng-ssw	7.6	15.5	38.9	4.4
avg	16.55	22.09	45.39	3.94
xho-zul	8.5	18	41.4	1.9
zul-sna	8.5	15	38.7	1.7
sna-afr	12	15.1	38	3.9
afr-ssw	5.3	11.2	34.3	7.2
ssw-tsn	14.4	15.4	38.9	2.9
tsn-tso	13.2	15.1	38.7	2.1
tso-nso	13.1	12	36.6	5.8
nso-xho	6.6	13.7	36.9	4
avg	10.2	14.44	37.94	3.69
overall avg	17.35	21.31	43.7	5.06

Table 6: The performance of our final system on the shared task test set. Δ CHRF2++ is the difference between the best submission and our system.

5.2 Translation between LRLs

Table 5 shows the performance of the OBPE+back-translation model before and after continued training for translation between LRLs. The model's performance improved on both the initial language pairs (in table 4) and the new translation directions. Moreover, $sna \rightarrow afr$ and $afr \rightarrow ssw$ were improved using only synthetic data (see section 4.2.1). We ascribe the success in improving the model's performance in translating between English and LRLs to the balancing approach (see section 4.2.2), as we used real training data (not back-translated sentences) in the continued training.

5.3 Official Results

Table 6 shows the results provided by the shared task organizers for our system as evaluated on a hidden test set. The table also compares the best constrained submission for each translation direction and our system. Our model did not achieve the best performance in any direction. However, the teams whose models performed better all trained on all languages included in the shared task (not just Southern African languages).

We hypothesize that this is the main reason for the gap in performance between our system and the better performing ones, as those models could benefit from more training data and increased cross-lingual transfer. The fact that our model performs relatively worse when translating into English provides some evidence for this: the other systems could benefit learning to translate to English in many more translation directions and with much more data in total. Given our computational resources, it would have required a total training time of 106 days to cover all language directions in the shared task. Unfortunately this was not feasible in the time provided for the shared task. The findings paper for the shared task presents more details about other teams' submissions (Adelani et al., 2022).

6 Conclusion

We have presented our multilingual neural MT model for 8 Southern African languages. Until recently, it would not have been possible to train a multilingual model for these languages because of data scarcity. During model development we found the benefits of multilingual modelling to be especially great for the lowest-resourced languages. Our results show that overlap BPE, back-translation, and synthetic training data generation are all valuable techniques for low-resource MT. More generally, we find multilingual modelling to be a fruitful approach to Southern African MT. For future work we would like to investigate further approaches for training large multilingual models for low-resource languages with a limited compute budget.

References

David Ifeoluwa Adelani, Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Akshita Bhagia, Marta Costa-

Jussá, Jesse Dodge, Fahim Faisal, Christian Federmann, Natalia Fedorova, Francisco Guzmán, Sergey Koshelev, Jean Maillard, Vukosi Marivate, Jonathan Mbuya, Safiyyah Saleem, and Holger Schwenk. 2022. Findings of the WMT 2022 shared task on large-scale machine translation evaluation for african languages. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257, Florence, Italy. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Auguste Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara E. Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Javier Ortiz Suárez, Iroro Fred Onome'Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics (TACL)*.

Baohao Liao, Shahram Khadivi, and Sanjika Hewavitharana. 2021. Back-translation for large-scale multilingual machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 418–424, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Vaidehi Patil, Partha Talukdar, and Sunita Sarawagi. 2022. Overlap-based vocabulary generation improves cross-lingual transfer among related languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–233, Dublin, Ireland. Association for Computational Linguistics.

Jerin Philip, Alexandre Berard, Matthias Gallé, and Laurent Besacier. 2020. Monolingual adapters for zero-shot neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4465–4470, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Jian Yang, Yuwei Yin, Shuming Ma, Dongdong Zhang, Zhoujun Li, and Furu Wei. 2022. High-resource language-specific training for multilingual neural machine translation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4461–4467. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.