ANVITA-African: A Multilingual Neural Machine Translation System for African Languages

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul,
Prasanna Kumar KR, Chitra Viswanathan
pavanpankaj333@gmail.com, {jsbhavani.cair, biswajit.cair, prasanna.cair, chitrav.cair}@gov.in
Centre for Artificial Intelligence and Robotics, CV Raman Nagar, Bangalore, India

Abstract

This paper describes ANVITA African NMT system submitted by team ANVITA for WMT 2022 shared task on Large-Scale Machine Translation Evaluation for African Languages under the constrained translation track. The team participated in 24 African languages to English MT directions. For better handling of relatively low resource language pairs and effective transfer learning, models are trained in multilingual setting. Heuristic based corpus filtering is applied and it improved performance by 0.04-2.06 BLEU across 22 out of 24 African→English directions and also improved training time by 5x. Use of deep transformer with 24 layers of encoder and 6 layers of decoder significantly improved performance by 1.1-7.7 BLEU across all the 24 African→English directions compared to base transformer. For effective selection of source vocabulary in multilingual setting, joint and language wise vocabulary selection strategies are explored at the source side. Use of language wise vocabulary selection however did not consistently improve performance of low resource languages in comparison to joint vocabulary selection. Empirical results indicate that training using deep transformer with filtered corpora seems to be a better choice than using base transformer on the whole corpora both in terms of accuracy and training time.

1 Introduction

Africa is very rich in languages, and around 1200 to 2100 languages are spoken in African countries[1], 24 African languages and 100 language pairs were selected for the WMT22 Large-Scale Machine Translation Evaluation for African Languages shared task Adelani et al. (2022b). Selected 24 African languages include Afrikaans(afr), Amharic(amh), Chichewa(nya), Hausa(hau), Igbo(ibo), Kamba(kam), Kinyarawanda(kin), Lingala(lin), Luganda(lug), Luo(luo), Nigerian Fulfulde(fuv), Northern Sotho(nso), Oromo(orm), shona(sna), Somali(som), Swahili(swh), Swati(ssw), Setswana(tsn), Umbundu(umb), Wolof(wol), Xhosa(xho), Xitsonga(tso), Yoruba(yor) and Zulu(zul) and language pairs include African-English, selective African-French, and African-African pairs, where many of the pairs fall under the low resource category. In this task, organizers permitted two submissions, Best scoring submission is considered as Primary model and other one being the Contrastive model. This paper describes our submission to WMT 2022 Large-Scale Machine Translation Evaluation for African Languages shared task where we participated for translation of 24 African languages to English. We are not officially given a rank as we didn't participate in all African MT directions.

2 Related Work

Developing quality machine translation system for low resource languages still remains a major challenge and many of the world languages fall under this category. Some of the recent developments do show that multilingual NMT is a promising direction. In massively multilingual neural machine translation, the authors have shown to train a single model for translating 102 languages to and from English and the results outperformed the strong bilingual baseline MT system especially for low resource languages Johnson et al. (2017). However, it cannot be generalized to all high and medium resource languages. Gowda et al. (2021) built a multilingual neural machine translation system capable of translating from 500 source languages to English which includes medium, low and extremely low resource languages. Zhang et al. (2020a) improved zero-shot translation in multilingual neural machine translation by random back translation. Kudugunta et al. (2019) have shown that represen-

[1] https://en.wikipedia.org/wiki/Languages_of_Africa

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 1090–1097
Abu Dhabi, December 7–8, 2022. ©2022 Association for Computational Linguistics

tations of high resource and/or linguistically similar languages are more robust when fine-tuning on an arbitrary language pair, which is critical to determining how much cross-lingual transfer can be expected in a zero or few-shot setting.

Zhou et al. (2021) shown deep architectures for neural machine translation and post ensemble have shown improved results on machine translation tasks. Zhang et al. (2020b) presented language independent heuristics for filtering noisy pairs from parallel corpus. Yang et al. (2021) proposed progressive training, in which the MT system is trained from shallow to deep architectures - increasing number of encoder and decoder layers. However, major improvement is observed while increasing encoder layers. Adelani et al. (2022a) created novel African corpus for 16 African languages and fine-tuned on pre-trained large MT models. Fan et al. (2021) demonstrated massively multilingual machine translation by training a single model that can translate between any pair of 100 languages.

3 Datasets

We used all the parallel corpora provided by WMT 2022 organizer. Corpus contain existing OPUS repository Tiedemann (2012), WMT 2022 novel corpus[2] and comprises of sources such as wikimedia, CCMatrix, CCAligned, bible-uedin, GNOME, XLEnt, QED,KDE4, mozilla-I10n, SPC, TED2020, Tatoeba, ELRC_2922, OpenSubtitles, Ubuntu, LAVA corpus2, MAFAND-MT Adelani et al. (2022a), KenTrans Wanzare et al. (2022), Kencorpus McOnyango et al. (2022), WebCrawl-African[3] and huggingface (provided by organisers) etc. Tiedemann (2012). Combining all a total 140 Million parallel sentences for the 24 African-English language pairs are extracted.

Language wise statistics of corpus used in our system is listed in Table 1.

4 System Overview

ANVITA African MT system comprises of two major sub systems: Data preprocessing and Model training under different strategies and architectural configurations followed by evaluation.

4.1 Data Preprocessing

As part of data preprocessing, we removed potentially noisy sentence pairs using the heuristics presented in Data Filtering subsection. To handle rare words and out of vocabulary words in the corpus we tokenized the training data using sentencepiece Kudo and Richardson (2018).

4.1.1 Data Filtering

As most of the corpora is extracted by automated techniques, there are chances of presence of noisy sentence pairs in the corpus. As transformer is known to be sensitive to corpus noise Liu et al. (2019) rigorous filtering was performed on the corpus based on heuristics adopted from Li et al. (2019), Vegi et al. (2021) and Pinnis (2018). Details of the heuristics used are listed below.

- F0: Filter out sentence pair, in which either source or target sentence is empty.
- F1: Filter out sentence pair, in which either source or target sentence length greater than 800 characters.
- F2: Filter out sentence pair in which length of source and target sentence ratio is greater than 2.5.
- F3: Filter out sentence pair in which length of source and target sentence ratio is less than 0.4.
- F4: Filter out sentence pair, if source or target sentence contains word having length greater than 10.
- F5: Filter out sentence pair, if source and target sentences are equal.
- F6: Filter out sentence pair, if source or target sentence length is less than 4.

Corpus statistics after applying heuristics based filtering is given in Table 1. By applying heuristics, approximately 31% of total parallel sentences amounting to 44802801 are removed as they are potentially noisy pairs. Relative impact of each filter is also captured in Table 1. Heuristics chosen are language agnostic but there is always a room for corpus and language dependent heuristics, specifically the threshold values.

Experiments are carried out to observe the effect of data filtering (Configuration B vs Configuration A). Configuration A and B are discussed in detail in section 5.

[2] https://statmt.org/wmt22/large-scale-multilingual-translation-task.html

[3] https://github.com/pavanpankaj/Web-Crawl-African

Table 1 Statistics of training data before and after applying heuristic based filtering
%Filt is wrt previous filter and cumm %Filt is wrt Raw corpus

African↔English	Raw	F1-filt	%Filt	F1 +F2+F3	%Filt	F1+F2+F3+F4	%Filt	F1+F2+F3+F4+F5	%Filt	F1+F2+F3+F4+F5+F6	%Filt	cumm %Filt
afr-en	14357809	14331047	0.19%	14258195	0.005%	13675966	4.08%	13586470	0.65%	12128497	10.73%	15.5%
amh-en	1192934	1192625	0.026%	1142002	4.24%	1128660	1.16%	1115194	1.19%	946778	15.10%	20.6%
nya-en	1548650	1548650	0%	1529186	1.25%	1519738	0.61%	1519738	0%	1415637	6.84	8.5%
hau-en	9114633	9113895	0.008%	4164957	54.30%	3729275	10.46%	3712130	0.46%	3349586	9.76%	63.2%
ibo-en	519236	517737	0.29%	500379	3.35%	492926	1.49%	473208	4.0%	372787	21.22%	28.2%
kam-en	1656152	1656152	0%	1617111	2.35%	1616089	0.06%	1616088	6.18%	1452332	10.13%	12.3%
kin-en	9881964	9880973	0.010%	9715917	1.67%	9603289	1.15%	9603287	2.08%	8595328	10.49%	13.02%
lin-en	2890688	2890688	0%	2833279	1.98%	2826725	0.23%	2826725	0	2294855	18.81%	20.6%
lug-en	3478641	3476981	0.004%	3399032	2.24%	3356346	1.26%	3356345	1*%	2667772	20.51%	23.3%
luo-en	2767133	2767133	0%	2724060	1.55%	2719714	0.16%	2719714	0%	2339916	14.0%	15.4%
fuv-en	1376106	1376105	0%	1356236	1.44%	1349177	0.52%	1349172	0.0003%	1256816	6.84%	8.6%
nso-en	3087818	3087812	0%	3014807	2.36%	3009047	0.19%	3004799	0.14%	2284885	23.96%	26.00%
orm-en	2793892	2793892	0%	2738209	1.99%	2703241	1.28%	2703241	0%	2139879	20.84%	23.4%
sna-en	8933636	8933542	0%	8709596	2.51%	8625135	0.97%	8625118	0.0001%	7335877	14.95%	17.88%
som-en	1459349	1458307	0.0007%	1358266	6.86%	1336338	1.61%	1321903	1.08%	1084345	17.97%	25.6%
swh-en	32811268	32805580	0.0001%	32374856	0.013%	32154373	0.68%	32022095	0.4%	28152884	12.08%	14.2%
ssw-en	165712	165712	0%	154561	6.73%	152334	1.44%	152334	0	93832	38.40%	43.3%
tsn-en	5931529	5931529	0%	5667299	4.45%	5614356	0.93%	5614356	0	4257859	24.16%	28.2%
umb-en	302951	302951	0%	295177	2.57%	294655	0.18%	294654	0.0003%	247063	16.15%	18.44%
wol-en	208084	208073	13.09*%	204758	1.59*%	202100	1.3%	201928	0.08%	138994	31.17%	33.2%
xho-en	29326727	29326373	0%	9926807	66.15%	9795968	1.31%	9775666	0.20%	7552496	22.74%	74.24%
tso-en	638447	638382	0.0001%	620738	2.76%	619539	0.19%	619480	0.009%	511184	17.48%	19.9%
yor-en	1710752	1709669	0.0006%	1665254	2.59%	1651573	0.82%	1630170	1.29%	1471404	9.74%	13.9%
zul-en	4091851	4091355	0.0001%	3969983	2.97%	3928045	1.06%	3917179	0.28%	3352155	14.42%	18.1%
Total	140245962	140205163	0.002%	113940665	18.7%	112104609	1.6%	111760994	0.3%	95443161	14.6%	31.2%

4.1.2 Tagging of Source Sentences

As most of the African languages follow Latin script, so as to tag input sentences based on languages we have added special tokens at the source side similar to Vegi et al. (2021). Tokens are generated using special symbols of length 4. Special symbols are used to avoid overlapping of tags with language vocabularies.

4.2 Vocabulary Selection

We experimented with various configurations of source side sentencepiece subword vocabularies. However, for target side we fixed sentence piece subword vocabulary size to 16K for all the configurations.

Source side vocabulary estimation is done based on the work of Gowda and May (2020), where it is shown that for low resource languages optimal BLEU score is obtained for relatively smaller subword vocabulary of size between 4K to 6K. Also as most of the African languages follow Latin script, there are also chance of large vocabulary(subword) overlap among the languages.

1. Source side vocabulary is set to 100K, jointly for all 24 languages and used in Configurations A,B,C and D. Please refer to Section 5 for more details on Configurations.
2. Language wise 4K to 6K subword vocabulary based on language corpus size, where 6K is used for the languages having more than 1 million sentence pairs and 4K for languages having less than 1 million size. Though it is expected a total vocabulary of around 130K but we obtained 75K combined vocabulary as there are many common subword vocabulary among languages. This is used in Configuration E.
3. We experimented with increasing source side joint vocabulary from 100K to 144K in which 120K subword vocabulary for top 18 high resource languages and remaining 24K for the remaining 6 languages.

4.3 Model Training

ANVITA African MT system used base transformer, deep transformer, ensemble techniques and used fairseq framework for training Ott et al. (2019).

4.3.1 Base Transformer: 6x6

Training configuration follows base transformer similar to Vaswani et al. (2017) and used 6 encoder and 6 decoder layers. Base transformer model is trained on all corpora provided by the organizer except WebCrawl African corpora.

4.3.2 Deep Transformer: 24x6

Training used 24 encoder and 6 decoder layers for 10 epochs with batch size 10240, dropout 0.3, word embedding size of 1024, adam optimizer, update-freq 8, heads 8, encoder and decoder feed forward dimension of 4096, batch type tokens, warm-up steps 4000, learning rate $5e^{-4}$. Training configurations are adopted from Yang et al. (2021). Constrained and Primary models are trained on all corpora provided by Organizers except WebCrawl

African corpora.

4.3.3 Ensemble

We ensembled last two epochs of Deep Transformer 24x6 i.e 11,12 and this was our primary submission for the shared task.

5 Experimental Evaluation and Result Analysis

Experiments carried out for 6 distinct configurations to assess effect of filtering, deep transformer and strategies used for vocabulary selection.

5.1 Configurations

- Configuration A: Experiment is carried out for 10 epochs on Base Transformer architecture with 6 encoder and 6 decoder layers with out applying data filtering on all corpus provided by WMT 22 except WebCrawl African corpora.

- Configuration B: Experiment is carried out for 10 epochs on Base Transformer architecture with 6 encoder and 6 decoder layers with heuristic based filtering on all corpus provided by WMT 22 except WebCrawl African corpora.

- Configuration C: Experiment is carried out for 10 epochs on Deep Transformer architecture (as discussed in 4.2.2) with 24 encoder and 6 decoder layers with heuristic based filtering on all corpus provided by WMT 22 except WebCrawl African corpora.

- Configuration D: Experiment is carried out for 10 epochs on Deep Transformer architecture (as discussed in 4.2.2) with 24 encoder and 6 decoder layers with heuristic based filtering on all corpus provided by WMT 22 including WebCrawl African corpora.

- Configuration E: Experiment is carried out for 10 epochs on Deep Transformer architecture (as discussed in 4.2.2) with 24 encoder and 6 decoder layers with heuristic based filtering and language wise subword vocabulary(as discussed in 4.1.2) on all corpus provided by WMT 22 including WebCrawl African corpora.

- Configuration F: Configuration C is carried out for 2 more epochs (i.e. 11 and 12) and applied ensembling of last 2 epochs i.e. 11 and 12.

5.2 Results and Analysis

ANVITA African→English MT system was evaluated on standard Flores200 dataset Costa-jussà et al. (2022) and evaluation was also done by the organizer of Large-Scale Machine Translation Evaluation for African Languages task on blind test sets Adelani et al. (2022b). Results of both the experiments are given below Tables 2 and 3. Configuration-F is our primary submission and Configuration-C is our Contrastive submission to the WMT 2022 shared task on Large-Scale Machine Translation Evaluation for African Languages. Due to computational and time constraints we were not able to submit a model with WebCrawl African corpora as a primary/constrained submission. All the experiments carried out on Nvidia RTX 8000 48GB single GPU system. Training base transformer (6 × 6) without filtering and with filtering took approximately 400 hours and 80 hours respectively for 10 epochs. Remaining all experiments used deep transformer took around 290 hours for 10 epochs.

Table 2 shows the results obtained when experiments are carried out with configurations A,B,C,D,E, and F.

In the following subsections, key insights obtained using configurations A,B,C,D, and E are presented with respect to effect of filtering, deep transformer, and individual language wise subword vocabulary selections. However Configuration F is not compared against other configurations, as Configuration F is a replica of Configuration C with 12 epochs and did not use WebCrawl African corpora.

5.2.1 Effect of Filtering: Configuration A vs B

1. Heuristic based filtering has shown significant improvement on BLEU and CHRF2++ ranging from 0.04-2.06 and 0.23-1.55 respectively on all 22 out of 24 African → English language directions.

2. Reduced training time from 400 hours (Configuration A) to 80 hours (Configuration B).

3. Decrease in BLEU score and CHRF2++ for two languages namely Nigerian Fulfulde(fuv) and Wolof(wol).

Table 2 Results of African to English models on Flores200
Tran: Transformer, (n): refers to n epochs, prim: Primary model submitted to task,
Contras: Contrastive model submitted to task, ISV:Individual subword vocabulary,
WA:WebCrawl African(corpus submitted as part of the task)

	A: Tran 6 × 6 (10)		B :Tran 6 × 6 + filt(10)		C:Tran 24 × 6 + filt (10)(Contras)		D:Tran 24 × 6 + filt + WA(10)		E:Tran 24 × 6 + filt + WA + ISV (10)		F:Tran 24 × 6 + filt + ensem (Prim) (11,12)	
Afr↔En	BLEU	CHRF2++	BLEU	CHRF2++	BLEU	CHRF2++	BLEU	CHRF2++	BLEU	CHRF2++	BLEU	CHRF2++
afr	50.97	70.64	51.8	71.28	**55.8**	74.185	55.73	**74.21**	55.56	74.07	56.38	74.52
amh	16.45	40.01	17.29	41.14	**24.39**	48.80	24.17	**48.82**	24.45	48.98	24.78	49.46
nya	18.42	40.09	18.5	40.90	22.45	**48.79**	**22.66**	45.46	22.35	44.37	22.90	45.148
hau	21.42	43.74	22.3	45.09	27.92	49.95	28.04	**50.18**	**28.25**	50.06	28.97	50.70
ibo	15.48	37.15	15.9	38.81	20.62	44.07	21.25	**44.44**	**22.35**	44.37	21.79	44.93
kam	6.98	23.86	7.44	24.65	9.24	28.26	**9.49**	**28.33**	8.78	27.09	9.41	27.91
kin	19.90	42.38	21.7	43.5	25.97	48.01	**26.15**	**48.34**	25.48	47.38	25.83	48.11
lin	14.26	35.06	15.22	36.34	19.34	40.80	**19.56**	**41.2**	18.18	39.82	19.4	40.82
lug	12.10	31.99	13.13	33.37	15.93	37.09	**16.69**	**37.73**	16.44	37.48	16.30	37.37
luo	13.41	33.64	13.08	33.87	17.34	**38.51**	**16.96**	38.32	16.62	38.04	17.54	38.58
nso	23.68	44.71	25.6	46.96	33.30	53.77	**33.54**	**54.52**	33.22	53.96	34.02	54.36
fuv	5.13	20.99	4.5	19.72	5.62	21.91	**5.82**	**21.95**	5.12	19.95	5.71	21.54
orm	6.75	24.70	7.38	26.24	11.27	31.55	**12.13**	**33.57**	11.94	33.06	11.67	32.05
sna	19.94	42.68	19.98	42.98	23.68	46.429	23.57	**46.73**	**24.23**	46.17	24.25	46.61
som	13.75	34.76	13.96	35.01	18.01	40.02	**17.80**	**40.02**	17.37	39.55	18.07	40.22
swh	33.71	56.30	35.77	57.85	41.01	62.23	**41.19**	**62.32**	40.60	61.99	41.34	62.49
ssw	16.73	38.00	17.61	39.18	23.68	45.79	**25.34**	**47.27**	24.6	46.61	24.49	46.15
tsn	18.01	39.7	18.35	40.63	22.66	44.97	**23.08**	**45.96**	22.88	45.27	23.2	45.65
tso	19.02	39.80	19.38	40.64	24.32	45.85	21.72	44.33	**25.35**	**47.09**	24.5	46.04
umb	3.98	20.58	4.33	21.57	**5.74**	**24.35**	5.55	24.27	5.41	23.44	5.65	23.87
wol	5.64	23.02	4.93	21.75	8.71	27.10	8.43	27.01	**8.85**	**27.35**	8.71	27.17
xho	25.01	47.1	25.18	47.49	31.8	53.78	32.01	53.84	**33.47**	**54.97**	32.53	54.09
yor	11.01	31.14	12.20	32.54	15.3	37.12	15.39	37.20	**15.98**	**38.14**	15.58	37.45
zul	27.07	49.64	28.17	51.01	33.4	55.52	33.79	55.70	**34.68**	**56.06**	34.34	55.78

5.2.2 Effect of Deep Transformer: Configuration B vs C

1. Deep transformer architecture (Configuration C) has shown significant improvement on BLEU and CHRF2++ ranging 1.12-7.7 and 2.19-7.89 respectively on all 24 African → English language directions.

2. As expected, it increased training time from 80 hours (Configuration B) to 290 hours (Configuration C), but still less than base transformer training time without filtering.

5.2.3 Effect of inclusion of WebCrawl African: Configuration C vs D

1. Inclusion of Our corpora-3, WebCrawl African (Configuration D) has shown improvement on BLEU ranging 0.01-1.66 for 12 out of 15 African→English translation directions and even by +0.18-0.68 for the 4 out of 9 African→English translation directions. However there is a marginal decrease in remaining African→English directions.

2. Inclusion of Our corpora-3, WebCrawl African (Configuration D) has shown improvement on CHRF2++ ranging 0-1.48 on 19 African → English language directions, however there is a marginal decrease in remaining directions.

5.2.4 Effect of ISV (Individual Subword Vocabulary): Configuration D vs E

ISV (Configuration E) has shown significant improvement on few language directions, however there is a marginal decrease of BLEU and CHRF2++ in majority of the directions and specifically 17 and 19 out of 24 respectively.

It is observed that increase of source side joint vocabulary beyond 100K does not improve performance and in fact decrease in BLEU score is observed for majority of the languages. Also use of language wise vocabulary selection did not consistently improve performance of low resource languages in comparison to joint vocabulary selection.

5.3 Comparison With Available Models

To the best of our knowledge, results using a single multilingual model covering all the 24 African languages to English is not available. Often meaningful comparison becomes hard as not all the reported results use same test-set used here. Yang et al. (2021) trained NMT model for translating

Table 3 Results of African to English models on blind test set from Organizer
Tran: Transformer, (n): refers to n epochs, blind:refers to blind test set used by Organizer for evaluation, prim: primary model submitted to the task, Contras: Contrastive model submitted to the task,
* represents the languages where evaluation was not provided by the Organizer

Afr↔En	F: Tran 24 × 6 (11,12)+ensemble(Prim)			C:Tran 24 × 6 (10) (Contras)		
	BLEU	spBLEU	CHRF2++	BLEU	spBLEU	CHRF2++
afr	56.1	59	74.4	55.8	58.7	74.2
amh	24.8	26	48.5	24.1	25.2	47.8
nya	23.8	26.5	45.7	23.1	26.2	45.5
hau	30.3	32.6	51.7	28.8	31.3	50.9
ibo	24.8	27.1	47.2	23.6	25.8	46.2
kam	10.3	12.4	28.2	10.3	12.4	28.4
kin	27.7	29.2	48.9	27.4	28.9	48.8
lin*	-	-	-	-	-	-
lug	16.6	18.7	37.2	16.5	18.5	36.7
luo	17.9	19.9	38.3	17.6	19.5	37.9
fuv	6.2	8	21.9	6.1	7.9	22
nso	34.1	35.9	54.1	33.7	35.5	53.6
orm	11.9	12.6	31.8	11.2	12	31.5
sna	25.3	28	46.7	24.6	27.6	46.3
som	21	22.7	42	20.7	22.2	41.4
swh	40.6	42	61.3	40.4	41.7	61
ssw	25.9	27.9	46.7	25.5	27.5	46.2
tsn	26.2	28.2	47.7	25.4	27.5	47.1
umb	6.4	8.2	24.6	6.2	8.1	24.7
wol*	-	-	-	-	-	-
xho	30	32.4	51.6	29.8	32.4	51.6
tso	25.3	27.4	46.2	25.3	27.2	46
yor	16.3	18.4	37.5	15.8	17.9	37
zul	33.6	35.6	54.4	32.5	34.9	54

101 languages from any to any directions and 12 out of 24 translation directions part of our submission are in common. Comparison on FLORES shows our model produced an improved results for 7 out of 12 African→English directions namely {Hausa, Chichewa, Swahili, Xhosa, Yoruba, Zulu }→English. Emezue and Dossou (2021) trained many to many models for African languages and 5 out 24 African translation directions part of our submission are in common. Our model showed an improvement for all the common 5 African→English directions namely {Igbo, Kinyarwanda, Xhosa, Yoruba, Swahili}→English.

6 Conclusion

This paper describes our submission to WMT 2022 shared task on Large-Scale Machine Translation Evaluation for African Languages under the constrained translation track. We focused on 24 African languages to English MT directions. Multilingual model with deep transformer showed significant improvement in BLEU and CHRF2++ scores across all 24 African to English MT directions. Vocabulary size of 4K to 6K per language for estimating size of joint source vocabulary seems to be a good choice in a multilingual setup. Heuristic based filtering did improve the BLEU scores. However the biggest gain of filtering observed is in terms of training time speed up by 5x. Empirical results indicate that training using deep transformer with filtered corpora seems to be a better choice than using base transformer on the whole corpora both in terms of MT accuracy and training time.

Acknowledgements

The authors would like to thank Director, CAIR for his constant encouragement, guidance and enablement.

References

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022a. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

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. 2022b. 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.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Chris Chinenye Emezue and Bonaventure F. P. Dossou. 2021. MMTAfrica: Multilingual machine translation for African languages. In *Proceedings of the Sixth Conference on Machine Translation*, pages 398–411, Online. 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, et al. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(107):1–48.

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3955–3964, Online. Association for Computational Linguistics.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Bei Li, Yinqiao Li, Chen Xu, Ye Lin, Jiqiang Liu, Hui Liu, Ziyang Wang, Yuhao Zhang, Nuo Xu, Zeyang Wang, et al. 2019. The niutrans machine translation systems for wmt19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 257–266.

Hairong Liu, Mingbo Ma, Liang Huang, Hao Xiong, and Zhongjun He. 2019. Robust neural machine translation with joint textual and phonetic embedding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3044–3049, Florence, Italy. Association for Computational Linguistics.

Owen McOnyango, Florence Indede, Lilian D.A. Wanzare, Barack Wanjawa, Edward Ombui, and

Lawrence Muchemi. 2022. Kencorpus: Kenyan Languages Corpus.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Mārcis Pinnis. 2018. Tilde's parallel corpus filtering methods for wmt 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 939–945.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Pavanpankaj Vegi, J Sivabhavani, Biswajit Paul, Chitra Viswanathan, and Prasanna Kumar KR. 2021. Anvita machine translation system for wat 2021 multiindicmt shared task. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 244–249.

Lilian D.A Wanzare, Florence Indede, Owen McOnyango, Edward Ombui, Barack Wanjawa, and Lawrence Muchemi. 2022. KenTrans: A Parallel Corpora for Swahili and local Kenyan Languages.

Jian Yang, Shuming Ma, Haoyang Huang, Dongdong Zhang, Li Dong, Shaohan Huang, Alexandre Muzio, Saksham Singhal, Hany Hassan, Xia Song, and Furu Wei. 2021. Multilingual machine translation systems from Microsoft for WMT21 shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 446–455, Online. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020a. 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.

Yuhao Zhang, Ziyang Wang, Runzhe Cao, Binghao Wei, Weiqiao Shan, Shuhan Zhou, Abudurexiti Reheman, Tao Zhou, Xin Zeng, Laohu Wang, et al. 2020b. The niutrans machine translation systems for wmt20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 338–345.

Shuhan Zhou, Tao Zhou, Binghao Wei, Yingfeng Luo, Yongyu Mu, Zefan Zhou, Chenglong Wang, Xuanjun Zhou, Chuanhao Lv, Yi Jing, Laohu Wang, Jingnan Zhang, Canan Huang, Zhongxiang Yan, Chi Hu, Bei Li, Tong Xiao, and Jingbo Zhu. 2021. The NiuTrans machine translation systems for WMT21. In *Proceedings of the Sixth Conference on Machine Translation*, pages 265–272, Online. Association for Computational Linguistics.