WebCrawl African : A Multilingual Parallel Corpora for African Languages

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Abhinav Mishra, Prashant Banjare, Prasanna Kumar KR, Chitra Viswanathan
pavanpankaj333@gmail.com, {jsbhavani, biswajit, abhinavmishra}.cair@gov.in,
{prashantban408, prasanna, chitrav}.cair@gov.in
Centre for Artificial Intelligence and Robotics, CV Raman Nagar, Bangalore

Abstract

WebCrawl African is a mixed domain multilingual parallel corpora for a pool of African languages compiled by ANVITA machine translation team of Centre for Artificial Intelligence and Robotics Lab, primarily for accelerating research on low-resource and extremely low-resource machine translation and is part of the submission to WMT 2022 shared task on Large-Scale Machine Translation Evaluation for African Languages under the data track. The corpora is compiled through web data mining and comprises 695K parallel sentences spanning 74 different language pairs from English and 15 African languages, many of which fall under low and extremely low resource categories. As a measure of corpora usefulness, a MNMT model for 24 African languages to English is trained by combining WebCrawl African corpora with existing corpus and evaluation on FLORES200 shows that inclusion of WebCrawl African corpora could improve BLEU score by 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 which are not part of WebCrawl African corpora. WebCrawl African corpora includes more parallel sentences for many language pairs in comparison to OPUS public repository. This data description paper captures creation of corpora and results obtained along with datasheet. The WebCrawl African corpora is hosted on GitHub repository [1].

1 Introduction

Parallel corpus play a vital role in the progress of data driven machine translation research and development. Availability of parallel corpora is still a concern for a large collection of world languages. Africa alone is home to an estimated 1200 to 2100 spoken languages[2] and more than 34 languages with 1 Million plus speakers. Many of these 34 languages and associated language pairs fall under the low and extremely low resource categories and machine translation researchers face setbacks due to unavailability of parallel corpus in public domain.

WebCrawl African corpora is a little step put forward towards addressing this issue. Languages covered in WebCrawl African corpora include (1) Afrikaans(afr), (2) Amharic(amh), (3) Chichewa(nya), (4) Hausa(hau), (5) Igbo(ibo), (6) Lingala(lin), (7) Luganda(lug), (8) Oromo(orm), (9) Swahili(swh), (10) Swati(ssw), (11) Tswana/Setswana(tsn), (12) Xhosa(xho), (13) Xitsonga(tso), (14) Yoruba(yor) (15) Zulu(zul) and (16) English and language pairs include African-English and African-African pairs. WebCrawl African is submitted as a part of Large-Scale Machine Translation Evaluation for African Languages shared task(data track) of WMT22 Adelani et al. (2022).

Rest of the paper is organized as follows. Section-2 briefly covers related work on parallel corpora compilation through web data mining. Section-3 covers content collection process followed for WebCrawl African corpora compilation, Section-4 details its alignment processes, Section-5 presents results and analysis. Finally Section-6 presents the datasheet capturing responses to bunch of critical questions capturing many relevant facets of WebCrawl African corpora ranging from motivation, composition, collection process, processing, users, distribution, maintenance and Section-7 conclusion.

2 Related Work

A good amount of translated text are available on the web. However compilation of parallel corpora from web which involves suitable source discovery, sentence extraction, sentence alignment and quality assessment, control is not trivial. Sentence

[1] https://github.com/pavanpankaj/Web-Crawl-African
[2] https://en.wikipedia.org/wiki/Languages_of_Africa

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 1076–1089
Abu Dhabi, December 7–8, 2022.

alignment is the most critical part and alignment techniques range from simple heuristics to neural sentence embedding. Bañón et al. (2020) compiled ParaCrawl corpora from selected websites comprising of 41 languages and Vec/Hun/BLEU-Aligned techniques were used for sentence alignment. Schwenk et al. (2021a) compiled WikiMatrix corpora from Wikipedia articles comprising of 85 languages and used cross-lingual LASER embeddings, distance based measures and FAISS library for fast sentence alignment. Schwenk et al. (2021b) created CCMatrix corpora from snapshots of CommonCrawl comprising of 137 languages and used cross-lingual LASER embeddings, distance based measures, FAISS library and vector compression for fast, storage efficient sentence alignment. Ramesh et al. (2022) compiled Samanantar corpora from selected websites comprising of 11 Indian languages and used LaBSE cross-lingual embeddings, cosine similarity and FAISS library for fast sentence alignment. Philip et al. (2021) proposed an iterative alignment-training-alignment method for expanding corpora of Indian languages.

3 Content Collection through Web Crawling

Creation of parallel corpora through web data mining, by making use of sources of multilingual translated text present on the web has almost became the de-facto technique for its cost effectiveness and scaling advantages. WebCrawl African corpora creation followed similar strategy. As a first step, search has been carried out to discover potential websites having the following characteristics.

- Source website preferably should comprise of large number of text articles published in more than one African languages or/and English.
- Source website should have permissive Copyright T&C and favourable content usage policy.
- Source website should aid in covering diverse information domains, writing styles, genre and contains text covering contemporary language usage etc.
- Source website should have reasonable credibility for ensuring content quality in terms how contents are populated, content review mechanism followed, chances of biases of various forms in hosted content etc.

We ended-up finding four websites namely (1) South African Government[3] comprising of Government communication, (2) Nalibali [4] comprising of multi-genre short stories, (3) Gotquestions[5] comprising of spiritual Q&A and (4) African gospel[6] comprising of song lyrics.

Text content is mined from these four identified websites following four step process.

- Analyze website layout and collect relevant content through suitable web crawler
- Preserve alignment supervision signals such as web-page/document level hyperlinks across languages etc, wherever available
- Extract plain text by stripping of html tags
- If script is latin then apply nltk English sentence tokenizer else manually check sentence delimiter and apply delimiter to tokenize sentences
- Further align sentences following alignment algorithms-1, 2, 3

A relative comparison of 4 websites in terms of their contributions to the WebCrawl African corpora is shown in Figure-1

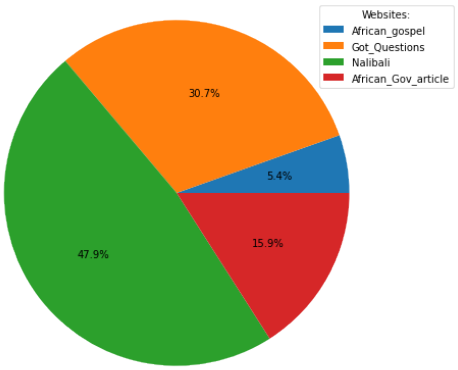

Figure 1: Source wise contributions in the WebCrawl African corpora

4 Alignment of Parallel Sentences

A good alignment strategy is expected to leverage alignment supervision signals available at the source websites. Since hyperlinks connecting

[3]https://www.gov.za/
[4]https://nalibali.org/
[5]https://www.gotquestions.org/
[6]https://africangospellyrics.com/

African and English language web-pages are available in the websites selected, the same is exploited for web-page level alignment and consequently search space for sentence alignment reduced significantly. Two different strategies are employed for sentence alignment duly leveraging the source websites information structure. Algorithm 1 is used for English-African parallel sentence alignment using cross-lingual embeddings and Algorithm 2 3 is used for fast African-African parallel sentence alignment based on common English sentences without using computationally expensive cross-lingual embeddings approach.

4.1 African-English Sentence Alignment

On an average, each web-page is having 200 to 250 sentences and hyperlinks to other language translated pages. Sentences from web-page aligned sources are extracted and segregated into source and target languages. Though web-pages are aligned, this unfortunately does not assure sentence level alignment due to improper sentence tokenization or even translation and format errors at the source. So a distinct need exists for carrying out sentence alignment exercise post segregation. Hence sentences are further aligned based on multilingual sentence encoders LASER[7] provided by the organizer of WMT22 Large-Scale Machine Translation Evaluation for African Languages shared task Adelani et al. (2022) and also heuristics. For a given row/sentence in source side, embeddings of all the target rows/sentences within a dynamic window around the source row is computed and the target row having maximum cosine similarity is selected as the source aligned sentence. Details are described in Algorithm 1. Time complexity of this African-English alignment algorithm depends on window-size. In worst case scenario, window-size can go up to number of source/target sentences and time complexity O(n^2), where n is max(#source sentences, #target sentences). Typically for the web-page aligned sources used, #source sentences or #target sentences range from 200 to 250.

4.2 African-African Sentence Alignment

The strategy employed for aligning African-African parallel sentences utilizes aligned African-English parallel sentences and does a fast alignment based on common English sentences without utilizing expensive cross-lingual embeddings.

[7]https://github.com/facebookresearch/LASER/tree/main/tasks/wmt22

5 Results and Analysis

We propose to evaluate the compiled WebCrawl African corpora in three ways. First, we present the distribution of extracted parallel sentences across language pairs. We then assess its usefulness by training a MNMT system for 24 African→English directions and finally compare it with resources available on public domain like OPUS.

5.1 Statistics of WebCrawl African Corpora

WebCrawl multilingual parallel corpora comprises a total of 695K mixed domain parallel sentences distributed non-uniformly over 74 language pairs. The parallel sentences are mined from webpages/documents such as government notifications, short stories, descriptive answers to spiritual questions and lyrics. The range of sentences varies from around 85 sentences (Hausa-Swati) to 64500 sentences (Swahili-English). For the monolingual corpora, the range varies from around 1,300 sentences for Igbo to 64,500 sentences for Swahili. Primary reason which influenced the number of parallel sentences is non-uniform coverage of text across languages on the websites sourced for the corpora compilation. Number of parallel sentences per language pair is captured in Table-1.

As per Table 1, African languages are relatively rich in vocabulary as compared to English. However this trend is not observed in case of Amharic, Hausa languages. Also its interesting to observe that even though Xhosa is not having the highest number of sentences, but has the highest vocabulary(xho-eng) among all pairs.

5.2 Usefulness of WebCrawl African Corpora

As a measure of corpora usefulness, two MNMT models for 24 African languages to English are trained. First one with the existing corpus and second one by combining WebCrawl African corpora with the existing corpus and both are evaluated on FLORES200 Costa-jussà et al. (2022). Results as shown in Table 2 show that inclusion of WebCrawl African corpora could improve BLEU score by 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 which are not part of WebCrawl African corpora, in spite being a tiny fraction as compared to the existing corpus. Potential reason could be WebCrawl African provides complimentary data to that of available

Algorithm 1 Algorithm for sentence alignment: African-English languages

Input: Tokenized sentences of srcLang(srcSentTok) and tgtLang(tgtSentTok), language encoder provided by organizers(src-lang-encoder),tgt-lang-encoder
Output: Aligned senteces for $srcSentTok$
$nsrc \leftarrow len(srcSentTok)$, $ntgt \leftarrow len(tgtSentTok)$
$windowSize \leftarrow abs(nsrc - ntgt) + 2$ ▷ abs: absolute value, 2 is added to windowSize as an additional margin to error i.e tokenization error, translator error
$i = 0$
while $i > nsrc$ **do**
    if $i - windowSize > 0$ and $i + windowSize < nsrc$ **then**
        $windowSent \leftarrow tgtSentTok[i - windowSize : i + windowSize]$
    else if $i - windowSize > 0$ and $i + windowSize >= ntgt$ **then**
        $windowSent \leftarrow tgtSentTok[i - windowSize : ntgt]$
    else if $i - windowSize < 0$ and $i + windowSize <= ntgt$ **then**
        $windowSent \leftarrow tgtSentTok[0 : i + windowSize]$
    else if $i - windowSize < 0$ and $i + windowSize >= ntgt$ **then**
        $windowSent \leftarrow tgtSentTok[0 : ntgt]$
    end if
    compute vector embedding of $srcEmbed[i] \leftarrow srcLangEncoder(i)$
    for all $j \in windowSent$ **do**
        compute vector embedding of $windowEmbed[j] \leftarrow tgtLangEncoder(j)$
        compute similarity $cosSimScore[j] \leftarrow cosine\_sim(srcEmbed[i], windowEmbed[j])$
    end for
    $maxind \leftarrow indexofmax(cosSimScore)$
    Required aligned sentence is $srcSentTok[i]$ with $windowSent[maxind]$
    $i = i + 1$
end while

Algorithm 2 Algorithm for sentence extraction/alignment: African-African languages

Input: parallel sentences of African-lang, English and Other-African-lang, English sentence pairs of all articles
Output: African-African-lang-p-sent, Other-African-African-lang-p-sent
j=0
while $j < len(articles)$ **do**
    sentence pairs in j^{th} article African-lang-en-p-sent,eng-p-sent and Other-African-lang-en-p-sent,eng-other-p-sent(afr-en pairs extracted from 1)
    Matching_indices = Compute_intersection(eng-p-sent, eng-other-p-sent)
    Align African-African-lang-p-sent, Other-African-African-lang-p-sent based on Matching_indices
    $African - African - lang - p - sent, Other - African - African - lang - p - sent$ are required gold parallel sentence pairs
    $j = j + 1$
end while

Algorithm 3 Algorithm for Compute_intersection

Input: Sentences of eng-p-sent, Sentences of eng-other-p-sent
edit_threshold = 4
Output: Returns list of tuples. for example [(1,1),(2,4)..] , means edit_distance(eng-p-sent[2], eng-other-p-sent[4]) <=edit_threshold
index1=0
index2=0
while $index1 < len(eng - p - sent)$ **do**
    while $index2 < len(eng - other - p - sent)$ **do**
        if edit_distance($eng-p-sent[index1], eng-other-p-sent[index2]) < edit\_threshold$ **then**
            tuple1 = $(index1, index2)$
            Matching_indices.append(tuple1)
        end if
        $index2 = index2 + 1$
    end while
    $index1 = index1 + 1$
end while
Return Matching_indices

in the existing corpus. Both the experiments used identical parameters and corpora used are only the only difference. For training, both WMT22 and WebCrawlAfrican+WMT22 corpus are further filtered using heuristics: (1) either source or target sentence is empty, (2) either source or target sentence length greater than 800 characters, (3) length of source and target sentence ratio is greater than 2.5 or length of source and target sentence ratio is less than 0.4 and (4) source or target sentence contains word having length greater than 10, (5) source or target sentence length is less than 4 and (6) source and target sentences are equal. Transformer with 24 layers of encoder and 6 layers of decoder are used for training both the models.

5.3 Comparison of WebCrawl African with OPUS Repository

A large part of African languages fall under the low and extremely low resource categories and do not have availability of parallel corpus of reasonable size in the public domain. A comparison of WebCrawl African corpora is carried out with the publicly available African parallel corpus listed on OPUS[8] repository in terms of parallel sentences.

Comparison results as captured in Figure-2 shows that out of 15 African-English language pairs compared, WebCrawl-African corpora has more number of parallel sentences for 7 African-English language pairs namely Chichewa-English, Lingala-English, Luganda-English, Oromo-English, Swati-English, Tswana-English, Tsonga-English languages as compared to OPUS public repository at the time of writing this paper. In fact WebCrawl-African corpora has 4 languages namely Chichewa, Luganda, Swati, and Tswana for which OPUS repository doesn't have even a single parallel corpora with any languages. Same goes for a few other African-African language pairs as well.

5.4 Corpora Quality

Though parallel corpora using web data mining approach can be created at scale, controlling quality of such corpora throws a major challenge. Noises ranging from source side errors such as incorrect translation, misspelling, incorrect grammar, biases of various forms and processing errors such as improper sentence tokenization, sentence alignment, additions, deletions etc often are of concern. In case of WebCrawl African corpora, the first choice made is to source content from credible websites, where website content is mostly generated in a controlled manner and contents are further reviewed. Also since the sources used have aligned web-pages so extracted sentence qualities are expected to be relatively better.

However, the authors could not analyze the corpora for translation correctness, biases and other quality metrics due to lack of knowledge on African

[8] https://opus.nlpl.eu/

Table 1: Statistics of WebCrawl African Parallel Corpora. (a,b,c) values in each box represents: $a = sentence\_count * 1000$, $b = unique\_source\_tokens * 1000$ and $c = unique\_target\_tokens * 1000$)

SrcLang(↓), TgtLang(→)	afr	amh	nya	eng	hau	ibo	lin	lug	orm	tsn	swh	ssw	xho	tso	yor	zul
Afrikaans (afr)	-	2.537 4.875 0.877	0.955 2.492 3.956	62.2 41.956 30.936	2.591 4.663 4.545	0.155 0.613 0.824	0	2.068 4.582 7.848	3.338 5.599 9.669	18.753 22.104 19.433	13.680 14.612 22.979	10.994 18.468 41.630	33.465 27.243 72.769	19.681 22.327 19.116	3.071 5.157 5.120	32.813 26.647 68.533
Amharic (amh)	2.537 0.877 4.875	-	0.634 0.212 3.105	4.6 1.294 6.737	2.562 0.891 4.781	0.117 0.012 0.742	0	1.65 0.760 6.744	2.816 1.065 9.264	0	3.130 1.240 9.361	0.091 0.012 0.996	0	0	2.792 1.076 5.401	0
Chichewa (nya)	0.955 3.956 2.492	0.634 3.105 0.212	-	1.4 5.180 3.177	0.92 3.912 2.584	0.16 0.922 0.837	0	0.987 4.478 4.484	0.947 3.908 4.027	0	0.964 3.959 3.474	0.136 0.820 1.149	0	0	0.92 3.926 1.498	0
English (eng)	62.2 30.936 41.956	4.6 6.737 1.294	1.4 3.177 5.18	-	5.6 7.48 6.606	1.1 1.45 2.219	1.1 0.956 1.53	3.6 5.478 10.71	7.0 8.164 14.252	25.9 20.191 22.946	64.5 27.103 59.569	14.4 16.428 47.934	46.2 24.481 85.768	24.4 19.158 21.254	6.3 7.585 7.647	50.9 25.022 84.648
Hausa (hau)	2.591 4.545 4.663	2.562 4.781 0.891	0.920 2.584 3.912	5.6 6.606 7.480	-	0.122 0.729 0.727	0	2.175 4.623 8.060	3.896 5.603 10.394	0	3.747 5.574 10.006	0.085 0.613 0.935	0	0	4.152 5.943 6.444	0
Igbo (ibo)	0.155 0.824 0.613	0.117 0.742 0.012	0.169 0.837 0.922	1.1 2.219 1.450	0.122 0.727 0.729	-	0	0.168 0.861 0.955	0.161 0.805 1.003	0	0.174 0.854 0.864	0.119 0.694 1.084	0	0	0.142 0.797 0.500	0
Lingala (lin)	0	0	0	1.1 1.53 0.956	0	0	-	0	0	0	0	0	0	0	0	0
Luganda (lug)	2.068 7.848 4.582	1.650 6.744 0.76	0.987 4.484 4.478	3.6 10.710 5.478	2.175 8.06 4.623	0.168 0.955 0.861	0	-	2.139 7.632 7.434	0	2.130 7.875 6.707	0.116 0.797 1.141	0	0	2.266 8.235 4.507	0
Oroma (orm)	3.338 9.669 5.599	2.816 9.264 1.065	0.947 4.027 3.908	7.0 14.252 8.164	3.896 10.394 5.603	0.161 1.003 0.805	0	2.139 7.434 7.632	-	0	4.583 11.654 11.437	0.123 0.820 1.069	0	0	4.333 10.966 6.477	0 25.022 84.648
Tswana/Setswana (tsn)	18.753 19.433 22.104	0	0	25.9 22.946 20.191	0	0	0	0	0	-	0	11.14 15.779 41.229	19.694 19.455 55.865	19.442 19.533 19.052	0	18.904 19.393 52.589
Swahili (swh)	13.68 22.979 14.612	3.13 9.361 1.24	0.964 3.474 3.959	64.5 59.569 27.103	3.747 10.006 5.574	0.174 0.864 0.854	0	2.13 6.707 7.875	4.583 11.437 11.654	0	-	0.133 0.737 1.194	0	0	4.134 10.725 6.309	0
Swati (ssw)	10.994 41.63 18.468	0.091 0.996 0.012	0.136 1.149 0.82	14.4 47.934 16.428	0.085 0.935 0.613	0.119 1.084 0.694	0	0.116 1.141 0.797	0.123 1.069 0.82	11.140 41.229 15.779	0.133 1.194 0.737	-	11.274 40.769 41.968	11.515 42.138 15.236	0.118 1.144 0.462	11.139 41.609 40.488
Xhosa (xho)	33.465 72.769 27.243	0	0	46.2 85.768 24.481	0	0	0	0	0	19.694 55.865 19.455	0	11.274 41.968 40.769	-	20.449 56.629 19.272	0	33.638 72.821 68.472
Xitsonga (tso)	19.681 19.116 22.327	0	0	24.4 21.254 19.158	0	0	0	0	0	19.442 19.052 19.533	0	11.515 15.236 42.138	20.449 19.272 56.629	-	0	20.342 19.390 53.702
Yoruba (yor)	3.071 5.12 5.157	2.792 5.401 1.076	0.920 1.498 3.926	6.3 7.647 7.585	4.152 6.444 5.943	0.142	0	2.266 4.507 8.235	4.333 6.477 10.966	0	4.134 6.309 10.725	0.118 0.462 1.144	0	0	-	0
Zulu (zul)	32.813 68.533 26.647	0	0	50.9 84.648 25.022	0	0	0	0	0	18.904 52.589 19.393	0	11.139 40.488 41.609	33.638 68.472 72.821	20.342 53.702 19.39	0	-
Total(sentences)	206.3	20.92	8.032	319.7	25.85	2.427	1.1	17.299	29.336	113.833	97.175	71.383	164.72	115.829	28.228	167.736

languages. Further human evaluation by language experts could not be carried out due to shortage of time and resources. As far as diversity of domains, genre, writing style and contemporary use of languages are concern, the source websites selected are expected to address them reasonably well. The details are covered in the datasheet presented in the next section.

6 Datasheets for WebCrawl African Corpora

Toward the growing consensus of having systematic dissipation of information on dataset to all its stakeholders by capturing all relevant facets, we followed Gebru et al. (2021) the idea of datasheets for datasets and further its adaptation by Costa-jussà et al. (2020) for MT taks. Datasheet for the WebCrawl African corpora is given below.

6.1 Motivation

(a) Who created the dataset(e.g., which team, research group) and on behalf of which entity (e.g. company, institution, organization)?

WebCrawl African corpora is compiled by ANVITA machine translation team of Centre for Artificial Intelligence and Robotics Lab based in Bangalore.

(b)Did they fund it themselves? If there is an associated grant, please provide the name of the grantor and the grant name and number

WebCrawl African corpora compilation work is fully supported by the Centre for Artificial Intelligence and Robotics Lab. No external grants are received or used for this work.

(c) For what purpose was the data set created?

African→English	WMT22* (#sentence)	WMT22*+ WebCrawlAfrican* (#sentence)	WMT22* (95M) (BLEU)	WMT22* (95M) (CHRF2++)	WMT22*+ WebCrawlAfrican*(260K) (BLEU)	WMT22*+ WebCrawlAfrican*(260K) (CHRF2++)
afr-en	12128497	12179628	**55.8**	74.185	55.73	**74.21**
amh-en	946778	950103	**24.39**	48.80	24.17	**48.82**
nya-en	1415637	1417004	22.45	**48.79**	**22.66**	45.46
hau-en	3349586	3354753	27.92	49.95	**28.04**	**50.18**
ibo-en	372787	373452	20.62	44.07	**21.25**	**44.44**
kam-en	1452332	1452332	9.24	28.26	**9.49**	**28.33**
kin-en	8595328	8595328	25.97	48.01	**26.15**	**48.34**
lin-en	2294855	2295671	19.34	40.80	**19.56**	**41.2**
lug-en	2667772	2670662	15.93	37.09	**16.69**	**37.73**
luo-en	2339916	2339916	**17.34**	**38.51**	16.96	38.32
fuv-en	1256816	1256816	5.62	21.91	**5.82**	**21.95**
nso-en	2284885	2284885	33.30	53.77	**33.54**	**54.52**
orm-en	2139879	2145917	11.27	31.55	**12.13**	**33.57**
sna-en	7335877	7335877	**23.68**	46.43	23.57	**46.73**
som-en	1084345	1084345	**18.01**	**40.02**	17.80	40.02
swh-en	28152884	28208419	41.01	62.23	**41.19**	**62.32**
ssw-en	93532	105225	23.68	45.79	**25.34**	**47.27**
tsn-en	4257859	4278691	22.66	44.97	**23.08**	**45.96**
umb-en	247063	247063	**5.74**	**24.35**	5.55	24.27
wol-en	138994	138994	**8.71**	**27.10**	8.43	27.01
xho-en	7552496	7588334	31.8	53.78	**32.01**	**53.84**
tso-en	511184	531823	**24.32**	**45.85**	21.72	44.33
yor-en	1471404	1477092	15.38	37.12	**15.39**	**37.20**
zul-en	3352155	3355480	33.4	55.52	**33.79**	**55.70**

Table 2: MT performance (BLEU, CHRF2++) with and without WebCrawl African Corpora.[*] Filtered

Was there a specific task in mind? If so, please specify the result type (e.g. unit) to be expected

WebCrawl African corpora is created primarily for accelerating research on low resource and extremely low resource machine translation. This corpora is also part of the submission to WMT 2022 shared task on Large-Scale Machine Translation Evaluation for African Languages under data track.

(d) Could any of these uses, or their results, interfere with human will or communicate a false reality?

No such thing is communicated to the authors. However, as machine translation is not free from biases, errors and may fail to portray actual essence of the translation or portray false, unfair realities, so such things can not be ruled out for WebCrawl African corpora and its usage as well.

(e) What is the antiquity of the file? Provide, please, the current date. The first version of WebCrawl African corpora was released on 10 May 2022. There was no further release till the time of writing this response.

(f) Has there been any monetary profit from the creation of this dataset?

The dataset is created and released mainly to aid research in MT and hoping to be useful for other NLP research as well. It's not for any monetary profit in the past, present and future as well.

6.2 Composition

(a) Is there any synthetic data in the dataset? If so, in what percentage?

WebCrawl African corpora does not contain any synthetic data.

(b) Are there multiple types of instances or is there just one type? Please specify the type(s), e.g. Raw data, preprocessed, symbolic.

WebCrawl African corpora comprises 695K parallel sentences spanning 74 language pairs from 15 African languages and English. African languages covered include Afrikaans(afr), Lingala(lin), Swati(ssw), Amharic(amh), Luganda(lug), Tswana/Setswana(tsn), Chichewa(nya), Hausa(hau), Oroma(orm), Xhosa(xho), Igbo(ibo), Xitsonga(tso), Yoruba(yor), Swahili(swh), and Zulu(zul). Source and Target parallel sentences are part of two separate files having the following naming convention.

Source file : webcrawl-african-{src-lang}-{tgt-lang}.{src-lang}

Target file : webcrawl-african-{src-lang}-{tgt-

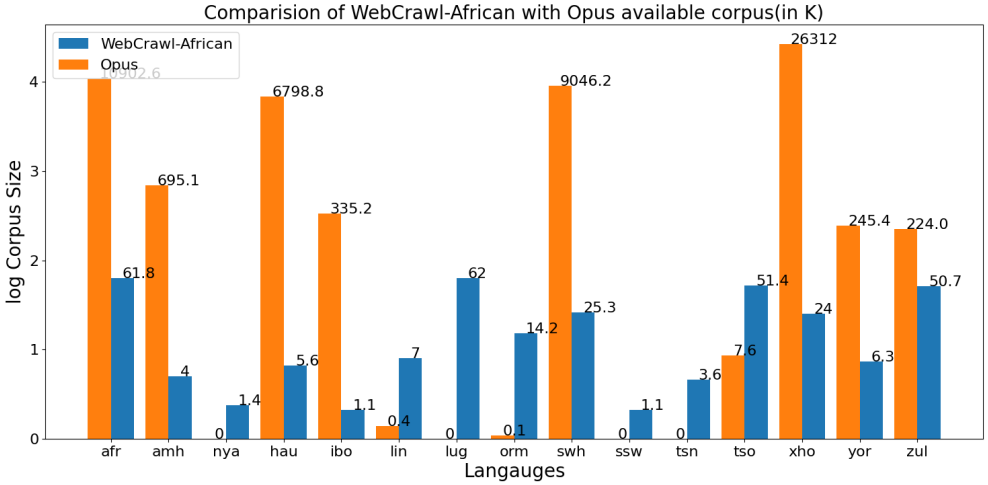

Figure 2: Comparison of WebCrawl-African corpora with the parallel corpus listed on OPUS repository

lang}.{tgt-lang}

src-lang and *tgt-lang* languages correspond to one of the 15 African languages and English part of WebCrawl African corpora and the whole corpora is spread over 148 files in 2 directories.

Monolingual corpora for language *src-lang* is available at *webcrawl-african-{src-lang}-eng.{src-lang}* file.

(c) What do the instances (of each type, if appropriate) that comprise the data set represent? (e.g. documents, photos, people, countries).

Instances represent parallel sentences aligned between two languages and stored in source and target files, following the naming convention mentioned above.

(d) How many instances (of each type, if appropriate) are there in total?

WebCrawl African parallel corpora comprises a total of 695K sentences (instances) distributed non-uniformly over 74 language pairs from 15 African languages and English. The range of sentences varies from around 85 sentences (Hausa-Swati) to 64,500 sentences (Swahili-English).

For the monolingual corpora available, the range of sentences varies from around 1,300 for Igbo to 64,500 sentences for Swahili. Complete count for each language pairs is available at the corpora hosting page `https://github.com/pavanpankaj/Web-Crawl-African`.

(e) Does the dataset contain all possible instances or is it just a sample of a larger set? i.e. Is the dataset different than an original one due to the preprocessing process? In case this dataset is a subset of another one, is the original dataset available?

WebCrawl African corpora is compiled by mining text from websites mentioned, through crawling and following sentence alignment techniques. Therefore, although the corpora is not a subset of any other corpora, it is limited by the text content crawled till the date of released of this corpora.

(f) Is there a label or a target associated with each of the instances? If so, please provide a description.

For any given language pair, a sentence in line number *i* and *Source language file : webcrawl-african-{src-lang}-{tgt-lang}.{src-lang}* will have an aligned target sentence in line number *i* and *Target language file : webcrawl-african-{src-lang}-{tgt-lang}.{tgt-lang}*. There are no other explicit labels associate with instances.

(g) What is the format of the data? e.g. .json, .xml, .csv .

For all language pairs, the aligned source and target sentences are kept in two seperate files following naming conventions mentioned in Section 6.2.(b) and all files are in UTF-8 plain text format.

(h) Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g. because it was unavailable). This does not include intentionally removed information, but might include, e.g. redacted text.

No such thing is reported. However, due to the automated techniques employed for corpora creation, some sentences may have missing words. Also there are language pairs for which no parallel sentences are present, for example Lingala does not have any language pairs with all other African languages included in WebCrawl African corpora.

(i) Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. Do not include missing information here.

No such thing is reported. However, due to the automated techniques employed for corpora creation, sentence misalignment error and misalignment induced noises in small proportion can not be ruled out. Additionally, content collected from African Gospel lyrics website where the content is generated through crowdsourcing with not so strict content review mechanism may have noises ranging from misspelling, grammatical errors and use of informal writings.

(j) Is there any verification that guarantees there is not institutionalization of unfair biases? Both regarding the dataset itself and the potential algorithms that could use it.

No such study is carried out or mechanism employed to assess and address corpora biases. Corpora is compiled by mining text from websites mentioned and inherited biases can not be ruled out. So both WebCrawl African corpora and translation algorithms could present biases.

(k) Are there recommended data splits, e.g. training, development/validation, testing? If so, please provide a description of these splits explaining the rationale behind them.

No specific splits are recommended.

(l) Is the dataset self-contained, or does it link to or otherwise rely on external resources? e.g., websites, tweets, other datasets. If it links to or relies on external resources, i) Are there any guarantees that they will exist, and remain constant over time? ii) Are there official archival versions of the complete dataset? i.e. including the external resources as they existed at the time the dataset was created. iii) Are there any restrictions (e.g. licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, if appropriate.

WebCrawl African corpora is self-contained and hosting page as mentioned contains the complete corpora.

(m) Does the dataset contain data that might be considered confidential? e.g. data that is protected by legal privilege or by doctor patient confidentiality, data that includes the content of individuals non-public communications. If so, please provide a description.

Corpora is compiled by mining text available in the public domain. So such a presence is unlikely.

(n) Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

Website content sourced for compiling corpora is meant for public consumption and genre includes government communication, short children stories, religious text and lyrics. So such anti-social content is unlikely. However, no review of the corpora is carried out from the perspective in question.

(o) Does the dataset relate to people? If so, please specify a) Whether the dataset identifies subpopulations or not. b) Whether the dataset identifies individual people or not. c) Whether it contains information that could vulnerate any individuals or their rights. c) Any other verified information on the topic that can be provided.

WebCrawl African corpora is compiled from open source content meant for public consumption and likely to reference people for the cause that made them appear publicly. Corpora does not include and express anything new which is not there in the public domain. However, no formal review of the corpora is carried out from the perspective in question.

(p) Does the dataset cover included languages equally?

Size of both parallel and monolingual corpora is not same for all the languages and language pairs included. Primary reason is the non-uniform coverage of text across languages on the websites sourced for the corpora compilation.

(q) Is there any evidence that the data may be somehow biased? i.e. towards gender, ethics, beliefs.

No study is carried out or mechanism employed to assess and address corpora biases. The corpora is compiled by mining text content available on the websites mentioned and inherited biases can not be ruled out.

(r) Is the data made up of formal text, informal text or both equitably?

WebCrawl African corpora comprises mostly formal text. However there are instances of informal content primarily coming from lyrics mined from African Gospel Lyrics website.

(s) Does the data contain incorrect language expressions on purpose? Does it contain slang terms? If that's the case, please provide which instances of the data correspond to these.

Given the genre of content hosted by the websites sourced for this corpora mining, such contents are unlikely. However, no review of the corpora is carried out from the perspective in question.

6.3 Collection Process

(a) Where was the data collected at? Please include as much detail; i.e. country, city, community, entity and so on.

WebCrawl African corpora is compiled by mining content hosted by websites (1) South African Government `https://www.gov.za/`, (2) Nalibali `https://nalibali.org/`, (3) Gotquestions `https://www.gotquestions.org/` and (4) African gospel `https://africangospellyrics.com/`. Websites comprise of text content covering government communication, multi-genre short stories, answers to spiritually related questions and gospel lyrics. A large part of it presumably written by the government officials, subject experts and volunteers primarily from the African countries and to some extent may be by the African speaking people from other countries. So data might be considered to have originated primarily from the African countries and other places around the globe as well. However, corpora compilation is carried out by the ANVITA team at Centre for Artificial Intelligence and Robotics, Bangalore.

(b) If the dataset is a sample from a larger set, what was the sampling strategy? i.e. deterministic, probabilistic with specific sampling probabilities.

WebCrawl African corpora is compiled by mining text content available on websites mentioned. The corpora is not a subset of any other corpora and no specific sampling was performed. However content is limited by the text crawled until the date of release of corpora.

(c) Are there any guarantees that the acquisition of the data did not violate any law or anyone's rights?

Websites having permissible copyright T&C and favourable content usage policy are used at the first place for content acquisition. Source websites permit usage and distribution of content for non-commercial, not-for-profit and fair use with due source acknowledgement. WebCrawl African corpora is hence released under CC-BY-NC-SA license for research purpose after intimation and with source acknowledgement. As long as WebCrawl African corpora license and source website copyright T&C and content usage policy is followed, one should safely assume that corpora acquisition and usage are unlikely to violate any laws or rights. Neither ANVITA team nor Centre for Artificial Intelligence and Robotics Lab holds any copyright over the WebCrawl African corpora. However, any derivatives of the corpora must acknowledge all sources including team ANVITA.

(d) Are there any guarantees that prove the data is reliable?

WebCrawl African corpora is created in an automated fashion without human verification like most of the large scale parallel corpora, thereby making it hard to guarantee provable reliability. However, as corpora is compiled by mining websites where content is mostly generated in a controlled manner and reviewed, makes the corpora reasonably reliable.

(e) Did the collection process involve the participation of individual people? If so, please report any information available regarding the following ques-

tions: Was the data collected from people directly? Did all the involved parts give their explicit consent? Is there any mechanism available to revoke this consent in the future, if desired?

As stated, content for the corpora compilation is directly sourced from websites mentioned and without direct participation of individual people.

(f) Has an analysis of the potential impact of the dataset and its use on data subjects been conducted? i.e. a data protection impact analysis. If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation. Neither such analysis is conducted nor any communication received on the subject.

(g) Were any ethical review processes conducted?

No ethical review processes were conducted.

(h) Does the data come from a single source or is it the result of a combination of data coming from different sources? In any case, please provide references.

WebCrawl African corpora is compiled by mining content hosted by four websites (i) South African Government `https://www.gov.za/`, (ii) Nalibali `https://nalibali.org/`, (iii) Gotquestions `https://www.gotquestions.org/` and (iv) African gospel `https://africangospellyrics.com/`. Websites comprise of text content covering government communication, multi-genre short stories, answers to spiritually related questions and gospel lyrics and contributed in creating mixed domain corpora. However, compilation is carried out by the ANVITA team.

(i) If the same content was to be collected from a different source, would it be similar?

It's likely to be similar if the content collected has a similar topic, genre, writing style and content distributions.

(j) Please specify any other information regarding the collection process. i.e. Who collected the data, whether they were compensated or not, what mechanisms were used. Please, only include if verified.

Content acquisition and corpora compilation is carried out by ANVITA team using a four steps process.

(i) Identification of websites based on content coverage, copyright T&C, usage policy and credibility.

(ii) Analysis of website layout and collection of relevant content through crawling preserving webpage/document level alignment signals, wherever available.

(iii) Extraction of plain text by stripping of html tags and splitting of text at sentence level

(iv) Alignment of parallel sentences across language pairs

6.4 Processing/Cleaning/Labelling

(a) Please specify any information regarding the preprocessing that you may know (e.g. the person who created the dataset has somehow explained it) or be able to find (e.g. there exists and informational site). Please, only include if verified. i.e. Was there any mechanism applied to obtain a neutral language? Were all instances preprocessed the same way?

WebCrawl African corpora is available as sentence aligned files. Preprocessing steps on raw crawled web-pages include striping off html tags, sentence tokenization and sentence alignment. Sentence alignment was carried out based on cross-lingual embeddings using LASER encoder and heuristics to a large extent. The entire corpora is preprocessed in the same way. No word/subword/character level tokenization or other pre-processing like filtering on parallel sentences were carried out. However, further filtering based on heuristics similar to the one used by the authors for training MT models may be carried out for better performance.

6.5 Users

(a) Has the dataset been used already? If so, please provide a description.

WebCrawl African corpora is used as a resource for the WMT 2022 shared task on Large-Scale Machine Translation Evaluation for African Languages Adelani et al. (2022). This corpora is also used by the ANVITA team for training MT model and results are presented in this paper.

(b) When was the dataset first released?

The initial release of WebCrawl African corpora was 10 May 2022.

(c) Is there a repository that links to any or all papers or systems that use this dataset? If so, please provide a link or any other access point.

WMT event is supposed to present a findings paper for the 2022 edition that may include such reference. Also corpora hosting page is likely to maintain such repository.

(d) What (other) tasks could the dataset be used for? Please include your own intentions, if any.

WebCrawl African corpora is primarily intended for machine translation tasks. It can also be used as monolingual corpora for tasks such as language modeling, corpus based language studies and few other NLP tasks with additional annotations.

(e) Are there tasks for which the dataset should not be used? If so, please provide a description.

There is no explicit task where this corpora should not be used. However, use of WebCrawl African corpora is not recommended as a benchmark corpora.

(f) Any other comments? i.e. Do the collection or preprocessing processes impact future uses?)

Like any large parallel corpora, WebCrawl African corpora is created in an automated fashion without human verification.

6.6 Distribution

(a) Please specify the source where you got the dataset from.

As mentioned, WebCrawl African corpora is compiled by mining text from web-pages hosted by (i) South African Government https://www.gov.za/, (ii) Nalibali https://nalibali.org/, (iii) Gotquestions https://www.gotquestions.org/ and (iv) African gospel https://africangospellyrics.com/

(b) When was the dataset first released?

WebCrawl African corpora was first released on 10 May 2022.

(c) Are there any restrictions regarding the distribution and/or usage of this data in any particular geographic regions?

No, there are no such restrictions.

(d) Is the dataset distributed under a copyright or other intellectual property (IP) license? And/or under applicable terms of use (ToU)? Please cite a verified source.

WebCrawl African Corpora distributed under CC-BY-NC-SA license. Barring commercial use, the license allows mostly unrestricted fair usage.

(e) Any other comments? i.e. How has the data been distributed? Who has access to the dataset? When was the dataset first distributed? Are there any other regulations on the dataset?

WebCrawl African Corpora is distributed through GitHub public hosting at https://github.com/pavanpankaj/Web-Crawl-African and also WMT 2022 website at https://www.statmt.org/wmt22/large-scale-multilingual-translation-task.html since 10 May 2022 under CC-BY-NC-SA license.

6.7 Maintenance

(a) Is there any verified manner of contacting the creator of the dataset?

All queries on WebCrawl African corpora should be sent to Pavan Pankaj Vegi at pavanpankaj333@gmail.com and Biswajit Paul at biswajit.cair@gov.in.

(b) Specify any limitations there might be to contributing to the dataset. i.e. Can anyone contribute to it? Can someone do it at all?

Scope exists for extending the corpora with additional parallel sentences and language pairs, specifically involving low and extremely low resource languages. Contribution can be done by contacting ANVITA team members at pavanpankaj333@gmail.com and biswajit.cair@gov.in

(c) Has any erratum been notified?

No erratum has been notified.

(d) Is there any verified information on whether the dataset will be updated in any form in the future? Is someone in charge of checking if any of the data has become irrelevant throughout time? If so, will it be removed or labeled somehow?

WebCrawl African corpora is likely to be updated with additional parallel sentences in future by the ANVITA team. Though chances of corpora becoming irrelevant in near future are less likely,

but if it happens, hosting page https://github.com/pavanpankaj/Web-Crawl-African will reflect the right status.

(e) Is there any available log about the changes performed previously in the dataset?

Not applicable, as the current version is the first version. However, log of future changes will be recorded at the corpora hosting page https://github.com/pavanpankaj/Web-Crawl-African

(f) Could changes to current legislation end the right-of-use of the dataset? WebCrawl African corpora is published under CC-BY-NC-SA license. We do not foresee any right-of-use changes in future.

(g) Any other comments? i.e. Is there someone supporting/hosting/maintaining the dataset? If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances?

Webcrawl African corpora is hosted at https://github.com/pavanpankaj/Web-Crawl-African and likely to be maintained by the ANVITA team..

7 Conclusion

This paper presented detailed description of WebCrawl African corpora. The paper also describes approach and design choices to systematically create parallel corpora and extend the WebCrawl African corpora through web data mining and alignment. WebCrawl African corpora compiled comprises 695K parallel sentences spanning 74 different language pairs from English and 15 African languages, many of which fall under low and extremely low resource categories. Webcrawl African corpora is hosted at https://github.com/pavanpankaj/Web-Crawl-African for non-commercial, not-for-profit and fair use. This corpora comprises sentences from multiple domains and includes government communication, short children stories, religious text and lyrics. Though human verification of the corpora was not carried out but favourable characteristics of selected source websites aided to address some of the quality concerns relatively better.

Experiments and evaluation of results show that inclusion of WebCrawl African corpora with WMT 2022 corpus has improved BLEU score by 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 which are not part of WebCrawl African corpora and also it has more parallel sentences for many language pairs in comparison to OPUS public repository.

WebCrawl African corpora is primarily intended for machine translation tasks, specially for accelerating research on low resource and extremely low resource machine translation. It can also be used as monolingual corpora for tasks such as language modeling, corpus based language studies and few other NLP tasks with additional annotations.

Acknowledgements

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

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 hared 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 Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Marta R Costa-jussà, Roger Creus, Oriol Domingo, Albert Domínguez, Miquel Escobar, Cayetana López, Marina Garcia, and Margarita Geleta. 2020. Mt-adapted datasheets for datasets: Template and repository. *arXiv preprint arXiv:2005.13156*.

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*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Jerin Philip, Shashank Siripragada, Vinay P Namboodiri, and CV Jawahar. 2021. Revisiting low resource status of indian languages in machine translation. In *8th ACM IKDD CODS and 26th COMAD*, pages 178–187.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.