Quality Does Matter: A Detailed Look at the Quality and Utility of Web-Mined Parallel Corpora

Surangika Ranathunga[1], Nisansa de Silva[2], Menan Velayuthan[2], Aloka Fernando[2], Charitha Rathnayake[2]
[1]Massey University, Palmerston North, New Zealand, 4443
[2]Dept. of Computer Science & Engineering, University of Moratuwa,10400, Sri Lanka
s.ranathunga@massey.ac.nz
{NisansaDdS,velayuthan.22,alokaf,charitha.18}@cse.mrt.ac.lk

Abstract

We conducted a detailed analysis on the quality of web-mined corpora for two low-resource languages (making three language pairs, English-Sinhala, English-Tamil and Sinhala-Tamil). We ranked each corpus according to a similarity measure and carried out an intrinsic and extrinsic evaluation on different portions of this ranked corpus. We show that there are significant quality differences between different portions of web-mined corpora and that the quality varies across languages and datasets. We also show that, for some web-mined datasets, Neural Machine Translation (NMT) models trained with their highest-ranked 25k portion can be on par with human-curated datasets.

1 Introduction

Despite the advances in NMT research, the availability of parallel corpora is still a deciding factor of NMT model performance. This puts low-resource languages at a clear disadvantage (Ranathunga et al., 2023). Even the use of Pre-trained Language Models (PLMs) is not quite enough to overcome the impact of data scarcity (Lee et al., 2022).

Publicly available web-mined parallel corpora (bitext) such as CCMatrix (Schwenk et al., 2021b), CCAlign (El-Kishky et al., 2020), WikiMatrix (Schwenk et al., 2021a), NLLB (Team et al., 2022), and ParaCrawl (Bañón et al., 2020) bring a glimmer of hope against this data scarcity problem. Compared to human-curated datasets, these are larger in quantity and contain data for hundreds of languages, including several low-resource languages. There are further initiatives to mine bitext for yet more languages as well (Bapna et al., 2022).

However, Kreutzer et al. (2022) analysed a sample of 100 sentence pairs from some of these corpora and showed that these web-mined corpora have serious quality issues, especially for low-resource languages. Lee et al. (2022) noticed a drop in NMT results when a model was trained using a random 100k sample of CCAlign. Khayrallah and Koehn (2018) injected different noise types found in web-mined corpora (by analysing a random sample) into a clean parallel corpus and showed that it has a debilitating impact on NMT performance.

These findings paint a grim picture of the utility of web-mined corpora. However, they all considered a random sample of these corpora to determine their quality. This implicitly assumes that the quality is consistent throughout the corpus.

In this research, we show that analysing a random sample of such large web-mined corpora can be misleading. We selected parallel corpora for two low-resource languages Sinhala and Tamil, which made three language pairs pairs: English-Sinhala (`En-Si`), English-Tamil (`En-Ta`) and Sinhala-Tamil (`Si-Ta`). Instead of quality checking a very small random sample of a web-mined corpus as done by Kreutzer et al. (2022), we ranked the sentence pairs by means of a similarity measure and extracted top 25k, bottom 25k and a random 25k portions of each corpus.

We improved the error taxonomy of Kreutzer et al. (2022) and carried out a human (intrinsic) evaluation on a random sample of 250 from each of these portions. Our results show that there are significant quality differences between the three portions, and the quality of the top 25k portion is much better than the other portions. We also noted major variations of quality across web-mined corpora belonging to different language pairs.

We then carried out an extrinsic evaluation. We separately trained NMT systems by using these top, bottom, as well as the random 25k samples of the corpora and tested them with two different evaluation sets. These results also showed that NMT models trained with the top 25k portion are significantly better. NMT models trained with the full version of some of these corpora were even lagging behind models trained with their top 25k portion. The NMT model trained with the top 25k

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 860–880

portion of the En-Si and En-Ta parts of the NLLB corpus performed even better than a model trained with a human-curated corpus.

We then fixed the translation issues in the top 25k of the NLLB corpus using human translators. The time taken to clean the corpus was slightly less than the time taken to translate the corpus from scratch. Although an NMT model trained with this cleaned corpus outperformed the uncleaned corpus, the resultant meagre gains cannot be justified when considering the time and money spent on the translators.

In summary, our results caution the researchers not to haphazardly use the web-mined corpora with just random sampling. Simply ranking a web-mined corpus first and then using only the high-quality portion would result in better accuracy in much less training time. We also hope other researchers (especially those working on low-resource languages) would carry out similar analyses for datasets of their languages. This will help future researchers make informed decisions when selecting web-mined corpora for NMT research.

2 Related Work

Web-mined parallel corpora are gathered from any available website without guarantees about quality. Khayrallah and Koehn (2018); Lee et al. (2022) pointed out that NMT systems built with such web-mined corpora have performance issues.

The common way to determine the quality of a parallel corpus is by analysing the performance of a Machine Translation system trained with that corpus (Khayrallah and Koehn, 2018; Schwenk et al., 2021a; Koehn et al., 2020). However, this does not indicate the types of noise in the corpus.

Human evaluation of the quality of parallel sentences (let them be web-mined, machine-generated, or human-generated) requires some criteria for the evaluators to make a judgement. Bojar et al. (2016) introduced the *Direct Assessment* criteria, where each sentence pair is ranked on a 0-100 scale. However, such a numerical scale does not shed light on the different types of noise in web-mined corpora.

Khayrallah and Koehn (2018) analysed a web-mined corpus and introduced the first categorisation of noise. The categories are: *misaligned sentences*, *mis-ordered words*, *wrong language*, *untranslated sentences*, and *short segments*. Herold et al. (2022) extended this categorisation with three new classes: *raw crawled data*, *over/under-translation*, and *synthetic translation*.

	CCAligned	Wikimatrix	CCMatrix	XLEnt	NLLB
En-Si	619,729	115,045	6,270,800	690,186	24,336,367
En-Ta	878,689	95,161	7,291,118	634,299	42,588,178
Si-Ta	-	-	215,965	153,532	1,493,318
Source	Common Crawl	Wikipedia	Common Crawl		
Filtering Level	document	sentence	sentence		
Alignment	LASER	LASER	LASER	LASER	LASER-3

Table 1: Dataset Statistics

In contrast to the above categorisations, Kreutzer et al. (2022)'s taxonomy has labels for both correct and erroneous sentence pairs: 1.) Correct translation - natural sentence, 2.) Correct translation but Boilerplate or low quality, 3.) Correct translation - short, 4.) Incorrect translation but both correct languages, 5.) Source OR target wrong language but both still linguistic content, and 6.) Not a language. Kreutzer et al. (2022) conducted a human evaluation using their taxonomy for three web-mined corpora (CCAligned, ParaCrawl v7.1, WikiMatrix) and covered data from both high and low resource languages. de Gibert Bonet et al. (2022) used that taxonomy to evaluate English-Catalan corpora.

3 Languages

We selected three language pairs: English-Sinhala (En-Si), English-Tamil (En-Ta) and Sinhala-Tamil (Si-Ta). Tamil (Ta) and Sinhala (Si) are large institutional languages (Eberhard et al., 2021). However, considering their data availability, Joshi et al. (2020) categorised Tamil as a mid-resource language and Sinhala as an extremely low-resource language. In the more recent language categorization by Ranathunga and de Silva (2022), Sinhala has moved one class up, and the position of Tamil is unchanged. Sinhala, in particular, is contained only in the island nation of Sri Lanka, and has only seen slow progress in language technologies (Ranathunga and de Silva, 2022; de Silva, 2023). But, being a multilingual country, translation systems are of utmost importance to Sri Lanka. This is particularly true for Si-Ta, as most government documents are first prepared in Sinhala and then translated to Tamil and English (Farhath et al., 2018).

4 Web-mined Parallel Corpora

Table 1 lists the web-mined corpora that we considered for evaluation. Other web-mined corpora available in OPUS (Tiedemann, 2012) were omitted because they did not have at least 100k samples for at least two of the language pairs we considered. Out

Error (E) Codes			
NL: Not a language: at least one of source and target are not linguistic content			
en	Many Melanesian societies, however, have become hostile towards same-sex relationships since the introduction of Christianity by European missionaries.[50]	si	[1]
en	Verily, you pass by them in the morning.	ta	37:137.
WL: Source OR target in some other language, but both still linguistic content			
en	Ի պատից պարոն Գոլջիի:	si	ගොල්ගි මහතා විසින් එතුමාගේ නමින් ම නම කෙරිණි.
en	I would probably go to Australia and I would study finance or communications.	si	Ben, sanırım Avustralya'ya gider ve finans ya da iletişim okurdum.
en	God is Sufficient (feat.	ta	ʒईश्वर पर्याप्त है (करतब)
UN: Most part of the source/target has been copied to target/source			
en	Create a new tab in an existing window rather than creating a new window	si	Create a new tab in an existing window rather than creating a new window
en	This certainly is the loss of revenue through Google AdSense.	ta	This certainly is the Google AdSense .
X: Correct source and target language, but the translation is completely wrong			
en	Several of William's children changed their surname as well.	si	එසේම සිය පුතායගේ නම ද වාශිෂ්ඨීපුත්‍ර පුලමාවි ලෙස වෙනස් කරයි.
en	"My lord would understand.	ta	நீங்கள் அறிந்தவரை இறைவன் என்பதன் எளிய அம்சங்களை விவரித்தால் என் போன்றவர்களுக்கு விளக்கம் கிடைக்கும், தடுமாற இயலாது.
Correct (C) Codes			
CS: Correct translation but very short sentences			
en	Supported platforms	si	සහාය දක්වන වෙදිකාවන්
en	Religion 101.	ta	மதம் 101
CB: Correct translation but boilerplate or low-quality. Requires considerable effort to derive the correct translation.			
en	No, you're right.	si	නැහැ ඔයා ඇත්ත කියන්නේ
en	It will be available for 30 days during which you can save, listen to, or share with others.	ta	இது 30 நாட்களுக்கு கிடைக்கும், இதன் போது நீங்கள் மற்றவர்களுடன் பகிர்ந்து கொள்ளலாம்.
CN: Near-perfect translation (minor grammar or spelling mistakes). Requires minor effort to derive the correct translation			
en	And in the Egyptian revolution, the Revolution 2.0, everyone has contributed something, small or big.	si	ඊවගේම ඊජිප්තියානු විප්ලවයේදී විප්ලවය අ0ක 2.0 සැවොම යම්කිසි අයුරකින් දායක වූවා
en	"50 children died yesterday."	ta	"நேற்று 50 குழந்தைகள் இறந்துவிட்டன.
CC: Perfect translation (no modification by the human is needed)			
en	A 5-year trusteeship was discussed, and a joint Soviet-American commission was established.	si	පස් අවුරුදු භාරකාරීත්වයක් පිළිබඳ සාකච්ඡා වූ අතර, ඒකාබද්ධ සෝවියට්-ඇමරිකන් කොමිසමක් ස්ථාපිත කෙරිණි.
en	It is our centuries-old traditional dance.	ta	இது எங்கள் நூற்றுண்டு பழமையான நடனம்.

Table 2: Quality Evaluation Taxonomy with En-Si & En-Ta examples

of the selected corpora, XLEnt (El-Kishky et al., 2021) was later omitted from human evaluation, because it has a significant amount of single-words or short phrases in its top 25k portion. These corpora are further described in Appendix A.

5 Quality Estimation by Humans

As mentioned earlier, Kreutzer et al. (2022) carried out the first human evaluation on the quality of web-mined corpora. Although they reported results for a large number of languages including low-resource languages, their discussion was mainly centred around the language-wise aggregated results. Thus they only used randomly selected 100 sentences from each language-specific corpus. de Gibert Bonet et al. (2022) carried out a similar study for Catalan-English, but they also considered only 100 samples from a corpus.

In contrast, we carried out a more detailed analysis of web-mined corpora belonging to the three language pairs by first ordering each parallel corpus according to the quality of the sentence pair. Our hypothesis is that the quality of a web-mined corpus is not consistent across a dataset, thus analysing a random portion of the corpus would not give a clear picture of the quality of the corpus.

Ours	Herold et al. (2022)'s
NL	
WL	Wrong Language (src|trg)
UN	Untranslated (src|trg)
X	Misaligned Sentences
CS	Short Segments (max. length)
CB	Misordered Words (src|trg), Raw Crawled Data, Over-/Under translation, Synthetic Translations
CN	
CC	

Table 3: Comparison of our taxonomy with the error Categories in Herold et al. (2022)

Participants: Fifteen translators were employed to conduct the human evaluation across the three language pairs. Evaluator selection and training details are in Appendix B.

Sample Selection: Calculating a similarity measure over the source and target sentence embeddings is a popular method to get an indication of the quality of a parallel sentence pair (Koehn et al., 2020). We picked LASER-3 (Heffernan et al., 2022) as our apparatus to score the alignment between the bitext. Heffernan et al. (2022) demonstrated that LASER-3 performs either on par or

Dataset		En-Si										En-Ta										Si-Ta									
		NL	WL	UN	X	E	CS	CB	CN	CC	C	NL	WL	UN	X	E	CS	CB	CN	CC	C	NL	WL	UN	X	E	CS	CB	CN	CC	C
CCAligned	Top	0.0	0.0	1.9	0.3	2.2	13.2	59.7	10.5	14.4	97.8	0.1	0.1	5.5	0.4	6.1	18.7	33.6	25.3	16.3	93.9	–	–	–	–	–	–	–	–	–	–
	Random	2.0	0.1	5.9	8.9	16.9	17.9	36.1	13.2	15.9	83.1	0.4	0.0	0.9	25.9	27.2	9.1	28.1	19.9	15.7	72.8	–	–	–	–	–	–	–	–	–	–
	Bottom	0.5	0.0	0.1	60.4	61.0	4.3	17.5	11.3	5.9	39.0	1.9	0.1	1.1	45.5	48.6	8.8	10.0	16.1	16.5	54.1	–	–	–	–	–	–	–	–	–	–
WikiMatrix	Top	0.3	0.1	2.1	16.3	18.8	6.1	40.8	12.0	22.3	81.2	0.9	6.3	15.9	46.1	69.2	1.6	10.3	7.5	11.5	30.9	–	–	–	–	–	–	–	–	–	–
	Random	0.3	0.1	0.0	86.1	86.5	1.2	7.9	2.9	1.5	13.5	0.7	0.9	1.2	91.9	94.7	0.3	4.1	0.7	0.3	5.4	–	–	–	–	–	–	–	–	–	–
	Bottom	0.0	2.7	0.3	88.5	91.5	1.2	6.9	0.4	0.0	8.5	1.3	5.2	0.8	88.7	96.0	0.3	2.1	1.2	0.4	4.0	–	–	–	–	–	–	–	–	–	–
CCMatrix	Top	0.0	0.0	7.1	0.1	7.2	8.7	37.5	14.3	32.4	92.9	0.0	0.0	51.5	3.6	55.1	2.7	27.5	8.5	6.3	45.0	0.1	5.5	0.5	2.1	8.2	9.3	26.4	34.3	21.7	91.7
	Random	0.0	0.0	1.6	31.3	32.9	6.1	27.6	22.5	10.8	67.0	0.1	0.0	2.9	83.5	86.5	0.3	8.5	3.1	1.6	13.5	0.0	2.1	0.8	31.3	34.2	0.9	34.7	23.2	6.9	65.7
	Bottom	0.0	1.3	0.8	27.2	29.3	7.3	47.3	8.5	7.5	70.7	0.0	0.1	0.0	83.1	83.2	0.0	10.1	4.3	2.4	16.8	0.0	1.2	0.1	50.1	51.4	1.1	31.7	11.3	4.4	48.6
NLLB	Top	0.0	0.5	0.4	19.1	20.0	6.5	36.0	18.8	18.7	80.0	0.0	0.4	0.3	11.1	11.8	0.4	21.6	25.2	41.1	88.3	0.0	0.0	0.3	1.9	2.2	0.3	22.3	40.5	34.8	97.9
	Random	0.1	0.4	0.7	54.5	55.7	1.3	27.5	10.5	4.9	44.2	0.1	0.0	0.5	43.3	1.9	43.9	31.6	10.9	11.6	98.0	56.0	0.3	0.0	20.0	20.3	1.2	44.5	22.3	11.7	79.7
	Bottom	0.0	0.0	1.9	56.9	58.8	4.7	27.1	8.1	1.3	41.2	0.0	0.0	0.0	51.9	51.9	1.6	28.9	11.1	6.5	48.1	0.0	0.0	0.1	34.7	34.8	0.0	42.0	20.3	2.9	65.2
NLLB (cleaned)	Translator 1	0.0	0.0	0.0	1.9	1.9	10.7	24.0	14.3	49.1	98.1	0.1	0.0	0.0	0.0	0.1	0.5	16.2	12.6	70.6	99.9	0.0	0.0	0.0	0.3	0.3	0.3	1.9	24.0	73.6	99.7
	Translator 2	0.0	0.1	0.0	1.9	2.0	7.2	21.3	13.0	55.6	98.0	0.0	0.0	0.0	0.1	0.1	0.8	16.9	10.1	72.1	99.9	0.0	0.0	0.0	0.4	0.4	0.3	4.3	38.0	57.0	99.6
	Translator 3	0.0	0.4	0.0	1.8	2.1	9.1	22.5	7.0	59.3	97.9	0.0	0.0	0.1	0.4	0.5	0.6	8.1	15.8	75.0	99.5	0.0	0.0	0.0	0.0	0.0	0.1	1.9	29.5	68.5	100.0

Table 4: The average percentage of tag counts over 3 independent evaluators for En-Si, En-Ta, and Si-Ta for 250 samples from top, bottom and random splits. C - sum of CS, CB, CN and CC. E - sum of NL, WL, UN, and X.

better than LaBSE[1], the other commonly used multilingual sentence encoder.

Sentences in each corpus were ordered by the LASER-3 score. For the NLLB corpus, we used the LASER-3 scores that were already provided within the dataset. For other datasets, we calculated this score[2]. From this sorted corpus, we randomly selected 250 sentences each from the top 25k split, the bottom 25k split, as well as from the entire corpus. There was no overlap between the sentences selected from the random set and the top/bottom sets. Once again, be reminded that Kreutzer et al. (2022) used only 100 random sentences from the entire corpus.

Taxonomy: Our error taxonomy shown in Table 2 is based on Kreutzer et al. (2022) and Herold et al. (2022). Unlike Kreutzer et al. (2022), we manually cleaned a web-mined corpus to determine its effect on NMT performance (see Section 9). Therefore our taxonomy indicates the level of human effort needed to fix the translation of a pair of sentences. We believe this provides more guidance to humans conducting quality evaluations of the corpora. Compared to Kreutzer et al. (2022), our taxonomy has two other differences: (1) We used WL to denote when the source or target is in some third language, and UN to denote when source or target has been copied to the other side. In contrast, Kreutzer et al. (2022) used WL to denote both of these scenarios. (2) Kreutzer et al. (2022) used CC to denote both perfect and near-perfect translations. In contrast, we used CC only for perfect translations and introduced CN for a near-perfect translation. While a bitext mining system may not be able to distinguish between CC and CN, this difference is important when manually cleaning the corpus.

Comparison of our taxonomy against Herold et al. (2022) is given in Table 3. Since they only focused on identifying errors, they do not have any category related to correct translation pairs. Herold et al. (2022) used their error categories to introduce synthetic errors to a clean corpus. Therefore they could easily generate data that corresponds to Mis-ordered Words (src|trg), Raw Crawled Data, Over/Under translation and Synthetic Translations. However, such errors are not directly distinguishable by a human. On the other hand, a sentence pair with at least one of these errors requires significant human effort to get cleaned. Therefore we grouped those categories as CB.

6 Human Evaluation Results

Each sentence pair was evaluated by three evaluators. The average agreement (measured in Pearson correlation) per language pair is as follows: En-Si 0.40, En-Ta 0.55 and Si-Ta 0.57 (Detailed results are in Table 9 of Appendix B). Results in Table 4 confirm 3 important points:

1. The quality of a web-mined corpus is not consistent throughout. We see drastic differences in quality between the top 25k and the bottom 25k. For example, the top 250 samples of the En-Si WikiMatrix corpus have 34.3% sentences falling into CC+CN categories, while its bottom portion has only 0.4% in the same categories.

2. Carrying out a human evaluation on a random sample as done by Kreutzer et al. (2022) portrays a high amount of quality issues. For WikiMatrix, CCMatrix, and NLLB, random sampling gives results that are closer to the bottom than the top. CCAligned defies this trend strongly in En-Si and weakly in En-Ta.

3. Quality of corpora can vary significantly depending on the language pair. For example, CCMatrix En-Si top 25k has 46.7% of CC+CN categories, and the same for En-Ta is 14.8%.

Together, these observations warn us against haphazardly using these web-mined corpora without

[1] https://tfhub.dev/google/LaBSE/2

[2] https://github.com/facebookresearch/LASER

studying their quality distribution. The result for non-English-centric Si-Ta is of particular interest. For Si-Ta, both NLLB and CCMatrix top portion seem to be extremely good. In fact, the 97.9% total value for the Correct (C) group is the highest across all the results.

Kreutzer et al. (2022) did not consider En-Ta or Si-Ta in their evaluation. Even for En-Si, only the ParaCrawl v7.1 corpus was considered. Therefore we cannot draw a direct comparison with their results. However, we can compare their micro-averaged results with our results for the random split, for the same corpora. For the CCAligned corpus, our random split results for both En-Si and En-Ta are significantly higher than Kreutzer et al. (2022)'s. In contrast, the same for WikiMatrix is lower than Kreutzer et al. (2022) by 10.24 and 18.34 (respectively). Even though Kreutzer et al. (2022) reported that 7.3% of the languages they analyzed did not contain a single correct sentence, we observed a similar phenomenon only with the bottom 25k split of WikiMatrix. These observations further justify the need for language-specific detailed analysis of web-mined corpora.

7 Qualitative Analysis of Corpora

In addition to the human evaluation discussed in the previous section, we also carried out a manual inspection of the top 25k portion of each corpus.

en	"What makes you think that it will be the truth, or even accurate?"
si	මහණෙනි, ඔබලා කුමකැයි සිතන්නාහු, රූපය නිත් යය හෝ වෙයිද?
	en: Monks, what do you think, is form constant?
en	And he opens up the refrigerator, and all he sees is the bright light.
ta	கதிரவன் தான் ஒளி யைத்தருகிறான், அனைத்தையும் காண செயகிறான்.
	en: The Sun is the one who gives light and makes everything visible.
en	God is All-knowing All-aware.
si	අපගේ ගාස්තෘ වූ බුදු රජාණන් වහන්සේ සියල්ල දන්නා සේක.
	en: Our teacher the Lord Buddha is all-knowing.
en	The two sea caves are linked, water goes in the one on the left and comes out the one on the right
ta	இரண்டு மகா கடல்கள் சங்கமிக்கும் பகுதி என்பதால், கடல் கொந்தளிப்பானது, இடதும் வலதுமாய், முன்னும் பின்னுமாய் கப்பலை அலைக் கழிக்கும்.
	en: As it is the confluence of two great oceans, sea turbulence, will toss the ship left and right, fore and back.
en	Is that evidence that he is God?
si	මෙවට කියන්නේ දෙයියෝ සාක්කි කියලද?
	en: Are these told as gods are the witnesses?
en	"Yes," they said. "you are not a person whom we doubt."
ta	"அவர்கள் சொல்வார்கள் ஃ "நீ எங்கள் தங்க மகனல்லவா!
	en: " "They will say, "Aren't you our golden son!

Table 5: Examples of *parallel* sentences from NLLB where the translated Si or Ta sentence has a different meaning than the original En sentences. We colour-coded the pairs of semantically close words that possibly contributed to the misalignment. Correct En translation of the Si/ Ta sentence is also given for comparison.

Similar to Kreutzer et al. (2022), in En-Si and En-Ta corpora, we found instances where sentences that are structurally and semantically similar but not parallel, presented as pairs. Table 5 shows some such interesting examples from NLLB (An extended version is in Appendix E as Table 14). We show instances where text from the *Bible* has been aligned with *Buddhist* scripture as well as instances where simple negation and noun matching have resulted in faulty alignments. Kreutzer et al. (2022) noted that such misaligned data may cause trained models to hallucinate fake *facts*.

Further, we observed some qualitative issues in the top 25k splits that are idiocentric to each dataset (or at least more prevalent in a particular dataset than others). CCMatrix has many untranslated/partially translated pairs. WikiMatrix, on the other hand, has many partial sentences. CCAligned has many concatenated lists coming from product advertisements (e.g., cameras, dongles, cables). Further, this dataset also has a comparatively higher amount of short entries. In general, NLLB was free of the above faults. However, as touched on in Table 5, the top pairs of NLLB are predominantly religious text with many misalignments. Informa-

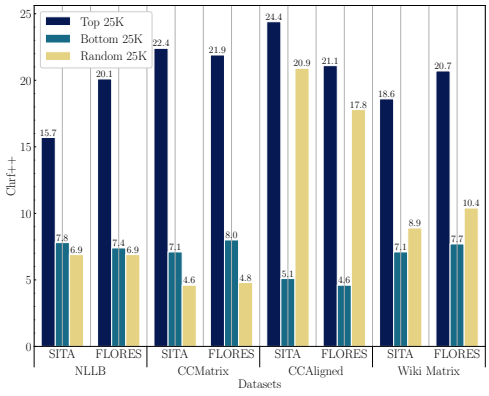

Figure 1: Vanilla-transformer performance trained on Top, Bottom and Random 25K splits of NLLB, CCMatrix, CCAligned and WikiMatrix for En-Si (higher the better).

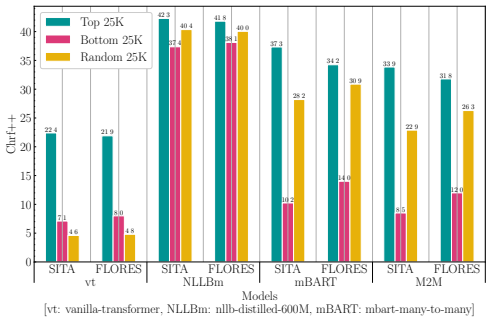

Figure 2: NMT results of different models trained on CCMatrix En-Si top, bottom and average 25K splits.

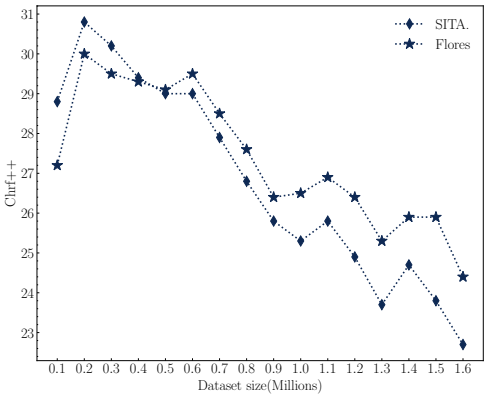

Figure 3: NMT results of vanilla transformer model trained on CCMatrix En-Si in jumps of 100K.

tion in some of the aligned NLLB sentences was not balanced (i.e. one side has more information).

The fact that NLLB has more religious text is worth noting because (1) NLLB is presented as a general domain dataset and not one in the religious domain (2) The phrasing and language used in these religious texts are more archaic than modern. Thus a model trained on the top 25k of NLLB might have a domain bias toward religious text and be unable to handle contemporary language.

8 Impact of Corpus Quality on NMT Model Performance

In Section 6, it was evident that different splits of a large web-mined corpus have different levels of quality. To determine whether this quality difference has any impact when it is used to train NMT models, we ran a series of experiments.

Dataset: For each corpus, we trained separate NMT models from the top, bottom, and random 25k portions of each of the web-mined En-Si corpora. We used two separate datasets for testing: FLORES-101 (Goyal et al., 2022), and the test set of the SITA parallel corpus (Fernando et al., 2020). FLORES was created from Wikipedia articles, and SITA from government documents of Sri Lanka.

Baseline Models: For Si-Ta, En-Si, and En-Ta, Thillainathan et al. (2021); Lee et al. (2022) showed that NMT models built on mBART (Tang et al., 2021) outperformed those built on vanilla Transformer models. NMT-specific models such as M2M (Fan et al., 2021) and NLLB (Team et al., 2022) (henceforth referred to as NLLBm, to distinguish from the NLLB dataset) have been shown to be generally better for low-resource languages (Zhu et al., 2023). However, these models have not been tested for the considered languages. Despite their performance, there is a possibility that the datasets considered in our experiments have already been included in these models (Jacovi et al., 2023). Thus, we trained vanilla Transformer NMT models with all data splits, and ran an ablation study with CCMatrix En-Si for NMT models trained on mBART, NLLB and M2M[3]. Model and training details are in Appendix D.

Results were recorded in chrF (Popović, 2015), chrF++ (Popović, 2017), BLEU (Papineni et al., 2002) and spBLEU (Goyal et al., 2022). chrF++ results are used in our discussion. All results are in Appendix F.

Results in Figure 1 (Raw result in Table 15 in Appendix F) confirm the observations we derived from human evaluation - the top 25k split is significantly better than the other splits. With respect to the SITA test set, the performance ordering of the corpora also tallies with human evaluation results for the Correct (C) category: CCAligned is the best, followed by CCMatrix, WikiMatrix, and NLLB. For FLORES test, CCMatrix is the best, followed by CCAligned, WikiMatrix, and NLLB. Interestingly, despite being created from Wikipedia, WikiMatrix could not beat CCAligned or CCMatrix for FLORES, which was also created from Wikipedia. The lowest result from NLLB could be due to its quality issues, as well as its religious content (see Section 7). Except in CCAligned, both bottom and random splits show roughly similar performance. The high result for the random split in CCAligned correlates with the higher value reported for the C category during human evaluation.

Figure 2 (Raw result in Table 16 in Appendix F) shows how NMT systems built with different pre-trained models perform on CCMatrix En-Si data splits. Overall, NMT models built on top of NLLBm show the best performance, followed by mBART and M2M-based models. Despite model-wise differences, these results reaffirm that NMT models trained with different splits of the same corpus have different levels of performance. This difference is least pronounced in the NLLBm model. Even in mBART and M2M models, the results gap between top and random splits is minimal, compared to the vanilla transformer model. This confirms that NMT systems built on pre-trained models

[3]mT5 (Xue et al., 2021) was not used as Nayak et al. (2023) showed that it lags behind mBART.

are more robust to noise in parallel corpora.

These findings naturally lead to the question '*what would happen to the NMT performance if the dataset size is gradually increased beyond 25k?*'. To answer this question, we trained vanilla transformer-based NMT models, by gradually increasing the size of the CCMatrix En-Si corpus up to 1.6M[4]. Figure 3 shows the results (Raw results are in Table 17 in Appendix F). Despite fluctuations, when the training dataset size increases, the results gradually decrease. Also note that for this corpus, the peak result is achieved when the training set is 200k. This number may vary from corpus to corpus[5].

We also trained vanilla Transformer models from the full CCMatrix, CCAligned and WikiMatrix for En-Si. Corresponding chrF++ results are 17.8, 41.7 and 17.3 (respectively) for SITA and 20.4, 31.7 and 19 (respectively) for FLORES. Comparing these values with those in Figure 2 shows that for some corpora, training an NMT model just with the top 25k split is better than using the full corpus.

9 Impact of Corpus Cleaning

9.1 Process and Human Evaluation

Creating high-quality corpora is a challenging task, especially for low-resource languages. In this context, employing human translators to clean web-mined corpora can be considered an alternative to creating parallel corpora from scratch.

In order to determine the benefit of corpus cleaning, we cleaned the top 25k of NLLB En-Si and En-Ta corpora. 11 En-Si translators and 16 En-Ta translators were used for this task[6]. Details of translator selection and training are shown in Table 10 of Appendix B. The translators were asked to first indicate the decision they took on a given sentence pair. The set of decisions and subsequent actions expected by the translator are given in Table 6.

Due to rewrites and deletes, the resulting corpus now has a final cleaned sentence pair count of 27,813 for En-Si and 26,526 for En-Ta. Table 7 shows the statistics of decisions taken by the translators. We see a significant number of updates, which confirms that the original corpus had more pairs falling into the C category. The lesser, but significant rewrites and very low count of deletes confirm that the E category was relatively small.

Recall that we conducted a human evaluation of 250 random samples from this portion of the NLLB corpus for both En-Si and En-Ta corpora. Each of these 250 samples was cleaned by three separate translators, while each sentence in the rest of the corpus was cleaned by a single translator. The 250 sentences of the top 25k portion of NLLB Si-Ta corpus that were used for quality estimation were also cleaned by three translators. The last three rows in Table 4 show this result. We see a significant drop in error (E) categories and a significant increase in CC category. However, human data cleaning has not produced a perfect result - had it been perfect, we should have seen 100% for CC+CS categories.

We manually reviewed the cleaned En-Si translation pairs that did not fall into CC or CS categories to identify why they were not cleaned to be perfect translation pairs. Our observations are as follows:

- NLLB has a high concentration of religious text. Jargon and structure used in the religious text are very different to contemporary vernacular. Some translators found it difficult to find equivalent wording in Sinhala for the religious-specific language.
- Some English sentences had structural issues. Some translators have not bothered to fix these structural issues and have simply translated that ill-formed English sentence into Sinhala.
- In the cases where the English sentence is partial (e.g. interrupted utterance), translating it to Sinhala was difficult due to differences in grammatical word ordering.
- Some English sentences that discuss ideas that are rooted in Western culture had no concise way of translating (e.g. I am taking her on a date).
- Spelling errors[7], errors caused due to overlooking punctuation errors.

Table 12 in Appendix C shows the time taken by translators for the corpus cleaning task. To produce 28,090 sentences from the noisy 25k En-Si corpus, the translators have collectively spent a total

[4]We did not go above 1.6M due to resource limitations.

[5]Thus, although we used 25k as our portion size, this number should not be taken as a universal cut-off value.

[6]In the case of the two translators who were involved in the corpus evaluation in addition to cleaning, we made sure not to (re)assign the samples they evaluated.

[7]Sinhala does not have a reliable spell corrector, and many errors can be easily overlooked (Sonnadara et al., 2021).

Sentence pair status	Decision	Subsequent action
Perfect translation (CC)	ACCEPT	Keep as it is
Acceptable translation, but En and/or Si has to be updated (CN, CS and CB in taxonomy)	UPDATE	Update En and/or Si
En AND Si both are either meaningless (i.e NL or WL), contain repetitive words (eg: No no no), or contain very short phrases (CS) (e.g. name of a place or a person)	DELETE	Keep as it is
En AND Si are meaningful sentences but not related (X)	REWRITE	Add two separate entries - En should be translated to Si, and Si should be translated to En
Only En OR Si are meaningful (i.e. one is NL, WL, UN, CS)	REWRITE	Rewrite the un-meaningful side to be the translation of the meaningful side

Table 6: Decision set employed for manual cleaning of the corpus (We remove ones marked as DELETE from the corpus before using it for NMT training.)

Decision	En-Si		En-Ta	
	Total Sentences	%	Total Sentences	%
Accept	4813	17.13	6621	24.70
Update	14852	52.87	15047	56.14
Re-write	8148	29.01	4858	18.13
Delete	277	0.99	275	1.03
Total	**28090**		**26801**	

Table 7: Summary of translator decisions

of 853:18(hr:m). On average, this is 1.8 minutes per sentence pair. To prepare a sample of a hundred sentence pairs, an average duration of 3hrs 3 minutes with a standard deviation of 1hr and 9 minutes was taken. Cleaning of 25k En-Ta sample produced 26,801 sentences consuming a total duration of 539:52 (hr:m). On average per sentence, the duration spent was 1.2 minutes. The average duration spent for a hundred sentences was 3hrs 47minutes with a standard deviation of 3hrs 57minutes. In both instances, the standard deviation is noticeably high. This is due to the individual capabilities/circumstances of translators or even a translator wrongly recording time(see Appendix C), it could also be due to the quality of the dataset portion received by a translator and the translator's judgment on the action to be carried out on a given sentence pair. This assumption is strengthened by results in Table 11 in Appendix C - there is a high variance in the actions selected by the translators.

To see if cleaning a web-mined corpus is more effective than translating a source from scratch, we selected three translators from the corpus cleaning task, gave them 100 sentences from NLLB En-Si corpus (that they had not seen before), and asked them to translate from scratch. We compared the time they took for the fresh translation and corpus cleaning.

As per Table 13 in Appendix C, corpus cleaning on average took 14 minutes less than fresh translation. However, the time taken to clean a corpus may vary depending on its quality. For the sake of completion, the freshly translated 100 sentences were evaluated by evaluators. CC, CN, and CB percentages are 57.00%, 10.67% and 32.33% respectively, which are on par with corpus cleaning results.

9.2 Impact on NMT Performance

For both En-Si and En-Ta, we built NMT models from the fully cleaned NLLB corpus, its top 25k, as well as from the random and the top 25k splits of SITA corpus. Figures 4 and 5 show the respective results. Raw results tables are in Table 18 in Appendix F[8]. For both language pairs, cleaned NLLB top 25k corpus beats the uncleaned version for SITA and FLORES test sets. But, compared to human effort to clean the corpus, this gain cannot be justified.

As per Figures 4 and 5, the top 25k split of both web-mined corpora performed better than SITA top 25k split for FLORES test set. Nayak et al. (2023) showed that NMT results could be affected by domain divergence. To determine whether this drop in SITA result is due to domain divergence, we calculated JS Divergence between different corpora (see Table 20 in Appendix G). The divergence between SITA and FLORES is the highest, but it is only slightly higher (0.1 points) than that of CCAligned. However, CCAligned result for FLORES is 2.3 chrF++ points higher than SITA. Therefore it is safe to assume the low performance of SITA may not be due to domain divergence, but due to its quality. However, the high domain divergence between NLLB and SITA is noteworthy. We remind

[8]Though we report result of SITA training set on SITA test set, this is misleading due to both coming from same corpus.

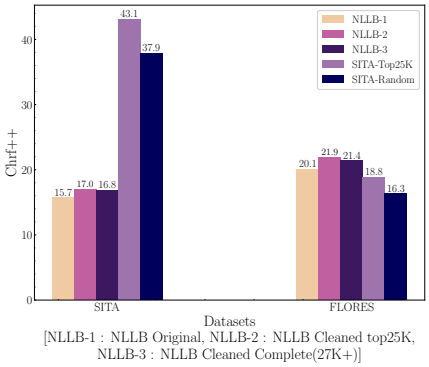

Figure 4: Vanilla transformer results for `En-Si` original NLLB Top 25K, NLLB cleaned Top 25K, NLLB cleaned full(27K+), SITA Top 25K, and SITA Random 25K.

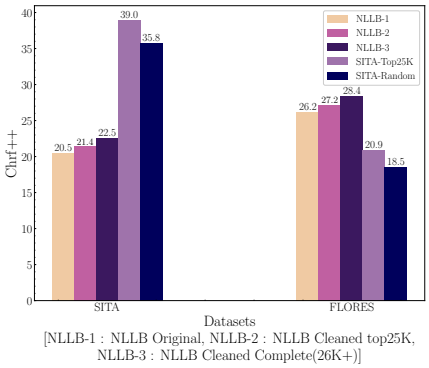

Figure 5: Vanilla transformer results for `En-Ta` original NLLB Top 25K, EnTa NLLB cleaned Top 25K, EnTa NLLB cleaned full(26K+), EnTa SITA Top 25K, and EnTa SITA Random 25K.

the reader that we noticed NLLB having higher amounts of religious content (see Section 7).

The full NLLB cleaned `En-Si` corpus of 27k+ lags behind the top 25k split. Similarly, for both language pairs, SITA random 25k lags behind the top 25k.

10 Impact of Embedding Technique

We used LASER-3 for ranking sentence pairs. The other commonly used measurements are LaBSE and XLM-R (Conneau et al., 2020). mBERT is another option, however, Sinhala is not included in this model. In fact, Fernando et al. (2023) compared LASER-1, LaBSE, and XLM-R embedding performance for sentence alignment and reported that LaBSE is superior to the other two. To determine whether the embedding technique has a noticeable impact, we ranked the CCMatrix `En-Si` corpus using LaBSE and XLM-R. Then we selected the top, bottom, and random splits from this

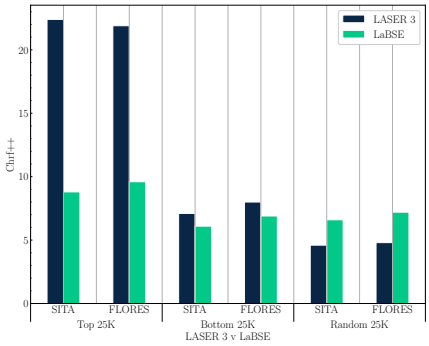

Figure 6: NMT Results on CCMatrix `En-Si` Top, Bottom and Random 25K for LASER-3 or LaBSE.

corpus and trained vanilla transformer-based NMT models. Figure 6 shows the comparison (Full result in Table 19 of Appendix F). XLM-R result is close to zero, so is not shown. Overall, the top 25k ranked by LASER-3 has a higher result than the other two. However, note that for a given language pair, the actual result may depend on the language representation in the model and the characteristics of the corpus. Thus, for a new pair of languages, it is worthwhile to experiment with different embedding models.

11 Conclusion

We presented a fine-grained evaluation of the quality of web-mined corpora for three low-resource language pairs. We showed that the quality of such corpora significantly varies across different portions. Our findings also indicate that simply using the highest quality portion of a web-mined corpus yields NMT results that may be on par with human-curated corpora in some instances. However, we are wary of further cleaning this top portion in hopes of better results, as the result gains do not justify the required human effort. Project artefacts are released and the details are shared in the project GitHub[9].

For our analysis, we considered the web-mined corpora without any pre-processing. If they were pre-processed (say) to remove duplicates, short phrases, or text in the wrong language, the performance of the embedding techniques may vary. We plan to investigate this in future. We also plan to expand this analysis to other low-resource languages.

[9] https://github.com/nlpcuom/quality-matters

Limitations

Our evaluation involves only three languages. This was inevitable because these are the only languages we had provisions to find human translators to carry out a meaningful evaluation. Due to financial constraints, we could carry out data cleaning only for the `En-Si` and `En-Ta` portions of the NLLB corpus. For the NLLB cleaning task, we reviewed only the first 100 sentences produced by the human translators. Therefore this corpus could still have some noise. From each corpus, we reviewed only 750 sentences. While this number is much larger than what Kreutzer et al. (2022) considered, it may still not be representative enough. Due to computing resource constraints, we could not train NMT models with all pre-trained models or train NMT models for various sizes of all parallel corpora. Our technique works only for languages included in embedding models such as LASER, LaBSE, and XLM-R.

Ethics Statement

We used publicly available parallel corpora that are free to use. Fernando et al. (2020) provided their dataset. We paid all the translators according to the government's stipulated rates. Before assigning them to the task, they were given a pilot to try out. They were given the chance to decide whether they were adequately compensated for their efforts. We only collected personal information that is needed for us to determine their suitability for the task and to arrange their payment. None of these personal details has been publicly released. More details are in the Appendix C. As mentioned under limitations, we could not manually review the corpus cleaned by translators. While they fixed the issues in a publicly available corpus, we cannot guarantee that the cleaned corpus does not have any unnecessary content that was not there in the original corpus.

Acknowledgements

This work was funded by the Google Award for Inclusion Research (AIR) 2022 received by Surangika Ranathunga and Nisansa de Silva.

References

Tahir Albayrak, Ram Herstein, Meltem Caber, Netanel Drori, Müjde Bideci, and Ron Berger. 2018. Exploring religious tourist experiences in jerusalem: The intersection of abrahamic religions. *Tourism Management*, 69:285–296.

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

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.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, et al. 2022. Building machine translation systems for the next thousand languages. *arXiv preprint arXiv:2205.03983*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ona de Gibert Bonet, Ksenia Kharitonova, Blanca Calvo Figueras, Jordi Armengol-Estapé, and Maite Melero. 2022. Quality versus quantity: Building Catalan-English MT resources. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 59–69, Marseille, France. European Language Resources Association.

Nisansa de Silva. 2023. Survey on Publicly Available Sinhala Natural Language Processing Tools and Research. *arXiv preprint arXiv:1906.02358v20*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2021. *Ethnologue: Languages of the World*. Dallas, Texas: SIL International.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Ahmed El-Kishky, Adithya Renduchintala, James Cross, Francisco Guzmán, and Philipp Koehn. 2021. XLEnt: Mining a large cross-lingual entity dataset with lexical-semantic-phonetic word alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10424–10430, Online and Punta Cana, Dominican Republic. 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. *The Journal of Machine Learning Research*, 22(1):4839–4886.

Fathima Farhath, Surangika Ranathunga, Sanath Jayasena, and Gihan Dias. 2018. Integration of bilingual lists for domain-specific statistical machine translation for sinhala-tamil. In *2018 Moratuwa Engineering Research Conference (MERCon)*, pages 538–543. IEEE.

Aloka Fernando, Surangika Ranathunga, and Gihan Dias. 2020. Data augmentation and terminology integration for domain-specific sinhala-english-tamil statistical machine translation. *arXiv preprint arXiv:2011.02821*.

Aloka Fernando, Surangika Ranathunga, Dilan Sachintha, Lakmali Piyarathna, and Charith Rajitha. 2023. Exploiting bilingual lexicons to improve multilingual embedding-based document and sentence alignment for low-resource languages. *Knowledge and Information Systems*, 65(2):571–612.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Christian Herold, Jan Rosendahl, Joris Vanvinckenroye, and Hermann Ney. 2022. Detecting various types of noise for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2542–2551, Dublin, Ireland. Association for Computational Linguistics.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. 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. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

En-Shiun Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Adelani, Ruisi Su, and Arya McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

Jinghui Lu, Maeve Henchion, and Brian Mac Namee. 2020. Diverging divergences: Examining variants of Jensen Shannon divergence for corpus comparison tasks. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6740–6744, Marseille, France. European Language Resources Association.

Shravan Nayak, Surangika Ranathunga, Sarubi Thillainathan, Rikki Hung, Anthony Rinaldi, Yining Wang, Jonah Mackey, Andrew Ho, and En-Shiun Annie Lee. 2023. Leveraging auxiliary domain parallel data in intermediate task fine-tuning for low-resource translation. *arXiv preprint arXiv:2306.01382*.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 823–848, Online only. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

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.

Charana Sonnadara, Surangika Ranathunga, and Sanath Jayasena. 2021. Sinhala spell correction: A novel benchmark with neural spell correction.

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. pages 3450–3466.

NLLB Team, 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*.

Sarubi Thillainathan, Surangika Ranathunga, and Sanath Jayasena. 2021. Fine-tuning self-supervised multilingual sequence-to-sequence models for extremely low-resource NMT. In *2021 Moratuwa Engineering Research Conference (MERCon)*, pages 432–437.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul,

Turkey. European Language Resources Association (ELRA).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

A Parallel Corpora used in the Study

All following artefacts were used consistent with their intended use when and where it was specified. The creators of the respective artefacts have checked whether their data contains any information that uniquely identifies individual people or offensive content. In the cases where the data is *updated* or *re-written* by translators as discussed in Section 8, the guideline discussed in Appendix C ensured that no information that uniquely identifies individual people or offensive content is inserted. The licences and terms of usage of the artefacts are as discussed in each of the cited sources below.

CCAligned (El-Kishky et al., 2020) is a dataset created using 68 snapshots of CommonCrawl[10]. Document alignment was done using FastText LangID (Joulin et al., 2016, 2017) by mapping documents with the same URL but different language codes. The alignments were then refined using LASER embeddings (Artetxe and Schwenk, 2019b).

WikiMatrix (Schwenk et al., 2021a) is a parallel corpus mined from Wikipedia. It has 135M parallel sentences in 1620 language pairs (85 languages). 34M of these are aligned with English. Duplicates have been removed after sentence splitting. FastText LangID has been used to identify the languages of the text and then LASER has been used to identify bitext.

CCMatrix (Schwenk et al., 2021b) was created using snapshots of CommonCrawl. It contains around 4.5 billion parallel sentences across 576 language pairs. In building CCMatrix, it was assumed the aligned sentence could appear anywhere on CommonCrawl. Thus, margin-based mining (Artetxe and Schwenk, 2019a) was used for sentence alignment.

NLLB (Team et al., 2022) was released along with a translation model of the same name. This dataset contains: (1) public primary bitext collected from various sources, (2) bitext mined with LASER-3 teacher-student training (Heffernan et al., 2022), and (3) Backtranslated bitext created from the monolingual corpus.

B Human Evaluator Details

Table 8 provides details about the human participants involved with the evaluation task. They all possess a minimum of one year of prior experience in translation. They are from Sri Lanka. We advertised for this work via social media. This set of translators was selected from a larger pool via a small test, by giving them ten sentences to translate.

Name	Experience (Years)	Qualification
En - Si		
Translator1	1	BA in German Language
Translator1	2	BA (Hons) Sinhala Sp.
Translator2	2	BA (Hons) in Translation Studies
Translator3	1	BA (Hons) in Translation Studies
Translator4	1	BA (Hons) in Translation Studies
Translator5	4	BA (Hons) in Translation Studies
Translator6	2	BA (Hons) in Translation Studies
En - Ta		
Translator7	7	BSc in Agriculture
Translator8	12	B.Sc. Applied Mathematics and Computing PGD in Professional Practice in English
Translator9	5	BSc (Hons) Engineering
Translator10	3	MBBS
Translator11	5	BA
Si - Ta		
Translator13	2	BA (Hons) in Translation Studies
Translator14	2	BA (Hons) in Translation Studies
Translator15	20	Diploma in Translation and Interpretation

Table 8: Details of Translators Involved in Corpus Evaluation task

A flow-chart (See Figure 7) was prepared to explain the evaluation task. Then they were given a pilot set to practice the task. We evaluated their work, refined the guidelines and provided them with the final instructions along with a demonstration video. They were paid for each sentence they evaluated. Before assigning work, we informed them of the rates. Thus, based on the time taken for the pilot task, the translators were given the option to decide whether they wanted to continue with the full task under the proposed payment rates. Table 9 contains the raw data used for the Pearson correlation study.

C Web-mined Corpus Cleaning

To clean the top 25k sentences from the NLLB corpus, translators were selected following the same procedure described in Section 5. Table 10 gives details about the human participants involved with the NLLB cleaning task.

Translators were issued a guideline (Figure 8) and a demonstration video. The authors reviewed the first 100 sentence pairs cleaned by the translators. Then an Extended Guidelines document was created to cover the common mistakes made during the task and to give specific instructions on the corrective action. The translators were asked

[10] http://commoncrawl.org/

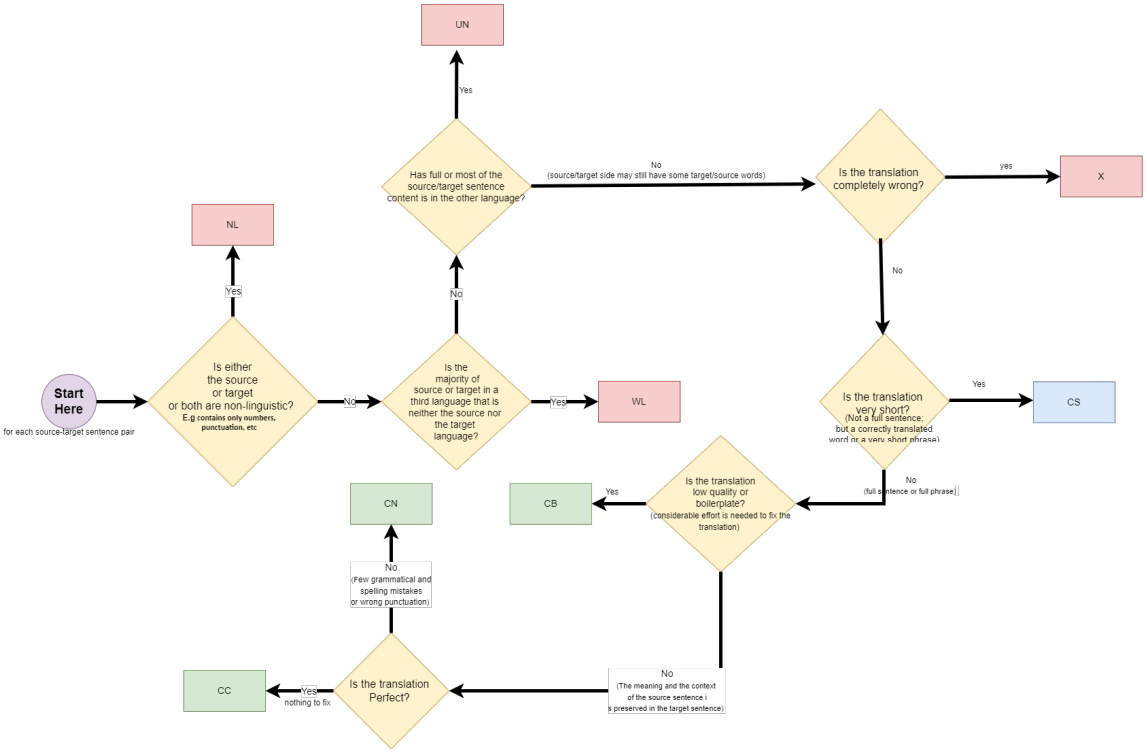

Figure 7: Flow-chart for Corpus Evaluation

Dataset			Eval1 v Eval2		Eval2 v Eval3		Eval1 v Eval3	
			Pearson-corr	p-value	Pearson-corr	p-value	Pearson-corr	p-value
CC Align	En-Si	Top	0.31	$4.22E-07$	-0.28	$5.79E-06$	-0.04	0.51
		Random	0.19	0.00	0.19	0.00	0.28	$8.85E-06$
		Bottom	0.78	$6.31E-52$	0.74	$2.61E-45$	0.68	$9.42E-36$
	En-Ta	Top	0.08	0.23	0.00	0.96	0.54	$1.41E-20$
		Random	0.55	$7.88E-21$	0.49	$7.61E-17$	0.58	$6.33E-24$
		Bottom	0.60	$1.99E-25$	0.72	$1.37E-41$	0.73	$2.21E-42$
Wikimatrix	En-Si	Top	0.19	0.00	0.27	$1.13E-05$	0.36	$6.36E-09$
		Random	0.34	$3.04E-08$	0.38	$7.43E-10$	0.29	$2.14E-06$
		Bottom	0.37	$1.06E-09$	0.52	$4.44E-19$	0.50	$2.22E-17$
	En-Ta	Top	0.65	$4.49E-31$	0.76	$1.10E-47$	0.68	$9.28E-36$
		Random	0.35	$1.42E-08$	0.17	0.01	0.37	$2.38E-09$
		Bottom	0.53	$1.34E-19$	0.56	$2.63E-22$	0.67	$9.30E-34$
CC Matrix	En-Si	Top	0.35	$1.26E-08$	0.52	$7.35E-19$	0.44	$2.28E-13$
		Random	0.43	$7.19E-13$	0.55	$1.91E-21$	0.37	$2.05E-09$
		Bottom	0.33	$1.16E-07$	1.00	0.00	0.33	$1.16E-07$
	En-Ta	Top	0.73	$1.66E-42$	0.75	$4.23E-47$	0.86	$9.73E-73$
		Random	0.59	$9.45E-25$	0.43	$1.06E-12$	0.49	$1.23E-16$
		Bottom	0.51	$8.11E-18$	0.42	$7.14E-12$	0.62	$6.13E-28$
	Si-Ta	Top	0.10	0.11	0.65	$1.04E-31$	0.16	0.03
		Random	0.63	$7.61E-29$	0.66	$1.32E-32$	0.55	$1.46E-21$
		Bottom	0.61	$2.50E-27$	0.74	$6.29E-45$	0.65	$3.89E-31$
NLLB	En-Si	Top	0.57	$2.80E-23$	0.51	$1.11E-17$	0.63	$1.39E-29$
		Random	0.38	$4.44E-10$	0.32	$2.80E-07$	0.47	$2.88E-15$
		Bottom	0.51	$4.31E-18$	0.46	$1.58E-14$	0.30	$1.38E-06$
	En-Ta	Top	0.64	$3.42E-30$	0.69	$4.23E-36$	0.68	$3.60E-35$
		Random	0.51	$1.01E-17$	0.45	$8.71E-14$	0.58	$1.89E-23$
		Bottom	0.57	$4.58E-23$	0.63	$2.19E-29$	0.69	$2.01E-36$
	Si-Ta	Top	0.56	$1.31E-21$	0.58	$1.15E-23$	0.64	$1.22E-30$
		Random	0.60	$3.85E-26$	0.66	$4.70E-32$	0.67	$2.61E-34$
		Bottom	0.54	$1.96E-20$	0.65	$2.59E-31$	0.66	$2.59E-32$

Table 9: Raw results used for Pearson correlation study for agreement between evaluators (Eval) on 250 samples for En-Si, En-Ta, and Si-Ta

to address the reviewer comments given for those hundred sentences. Once the reviewers were satisfied that a translator had fully understood the task, they were given the OK to continue with corpus cleaning. They were asked to record the exact time they spent on the corpus cleaning task.

Translators were paid as follows: For reading and deciding on the action to be carried out on a

Name	Experience (Years)	Qualification
En - Si		
Translator16	4	BA (Special) in Translation Studies
Translator17	1	BA (Hons) in Translation Studies
Translator18	1	BA (Hons) Business and Academic Chinese
Translator19	4	MBBS
Translator20	1	BA (Hons) in Translation Studies
Translator21	2	BA (Hons) in Translation Studies
Translator22	1	BEng (Hons) Technology (Mech Eng)
Translator23	3	BA (Special) in Translation Studies
Translator24	1	BA(Hons) in French Language and Literature
Translator25	2	BA (Hons.) in Translation Studies
Translator26	3	Certificate in Effective English
En - Ta		
Translator27	6	BSc (Hons) in Information Technology
Translator28	1	Bsc (Hons) Business Information System
Translator29	4	BA (Hons) in Translation Studies
Translator30	1	BSc Environmental Conservation and Management
Translator31	1	Bachelor of Unani Medicine and Surgery
Translator32	4	Bachelor of Information Technology
Translator33	1	BSc (Hons) Eng. sp. in Computer Science and Eng.
Translator34	1.5	BSc (Hons) Eng. sp. in Chemical and Process Eng.
Translator35	3	BSc (Hons) in Town and Country Planning
Translator36	2.5	B.Tech in Chemical Engineering
Translator37	1	BSc (Hons) in Nursing
Translator38	4	BA (Hons.) in Translation Studies
Translator39	8	Master of Business Management
Translator40	2	BA (Hons.) in Translation Studies
Si - Ta		
Translator41	6	BA (Hons) in Translation Studies
Translator42	1	BA in Social Sciences
Translator43	35	MA - Linguistic

Table 10: Snapshot of a translator's worksheet

sentence pair, a fixed amount was paid. When the translator updates or rewrites sentences, they are paid for each word they write/modify. They were informed of the rates in advance and were given the chance to opt-out of the task after participating in the pilot task. Table 11 shows the counts of decisions taken by each translator.

In Table 12, we have summarised the number of sentence pairs cleaned by each translator and the total time taken for it. Owing to the availability of translators, the number of sentence-pairs cleaned by each person was different. Therefore in our calculation, the average time spent by each translator to clean 100 sentence-pairs was considered. Based on these statistics, to clean a sample of hundred sentence pairs from the top 25k of the `En-Si` corpus, an average duration of 3hrs 3 minutes with a standard deviation of 1hr and 9 minutes was taken. For `En-Ta` the average duration was 3hrs 37minutes with a standard deviation of 3hrs 57minutes. We contacted the translators to confirm the times they reported. The three translators who have taken the longest time revealed that they have been recovering from illness/accident, therefore the work had been slow.

To compare these durations to what was taken for translating from scratch, we then asked three translators to provide fresh `Si` translations for a hundred `En` sentences and to record the time taken. Then we calculated and compared the average time taken along with the standard deviation with the values obtained for the corpus cleaning duration. This information is available in Table 13. The difference between the averages comes to 14 minutes, which means as the number of sentences increases, the time taken for the cleaning task will further be increased.

D Model Details

As discussed under Section 8, experiments were performed using three state-of-the-art (SOTA) NMT models (NLLB, mBART, and M2M) and vanilla transformer. To perform a fair evaluation between the SOTA NMT models, we chose model variants with similar model sizes. The model sizes utilized in our experiments for NLLB, mBART, and M2M are approximately 600 Million (M), 600M, and 418M, respectively. We perform bilingual fine-tuning on the SOTA models by training up to 3 epochs with a learning rate of 5×10^{-5}, maximum token length of 200 was set for both source and target. A batch-size of 10 was used for fine-tuning. We utilized the implementations provided by the HuggingFace Transformer library (Wolf et al., 2020), and Nvidia Quadro RTX 6000 for hardware-level parallelism. For the decoding process, default settings provided by HuggingFace were retained for each model. In the case of mBART and M2M, a beam search with a beam size of 5 was employed, while for NLLB, a beam size of 4 was utilized.

Vanilla-transformer: We train the Transformer models implemented in FAIRSEQ library (Ott et al., 2019) for our experiments. We train a model consisting of 6 encoder and decoder layers, encoder and decoder embedding of 256, 2 attention heads, dropout of 0.4, the learning rate of 1×10^{-7}, weight decay of 1×10^{-4} and a batch-size of 32. For decoding, we use beam search with a beam size of 5.

E Extended Misalignment Analysis

Table 14 is an extended version of Table 5. As mentioned in the discussion in Section 7, there are some interesting observations where text from the *Bible* has been aligned with *Buddhist* scripture. There were indeed multiple occurrences of *Bible* being aligned with the *Quran*. However, unlike in the case of *Bible-Buddhist* pairings, we are not showing *Bible-Quran* pairings here given that both of

Translator	Total Sentence-pairs	Accept		Update		Re-write		Delete	
		Decision Count	%	Decision Count	%	Decision Count	%	Decision Count	%
Translator16	652	29	4.45	540	82.82	74	11.35	9	1.38
Translator17	1523	414	27.18	664	43.60	403	26.46	42	2.76
Translator18	3402	367	10.79	1426	41.92	1608	47.27	1	0.03
Translator19	3636	261	7.18	1725	47.44	1647	45.30	3	0.08
Translator20	2288	920	40.21	1094	47.81	235	10.27	39	1.70
Translator21	2726	660	24.21	1272	46.66	785	28.80	9	0.33
Translator22	2669	594	22.26	1219	45.67	756	28.33	100	3.75
Translator23	2298	349	15.19	1452	63.19	458	19.93	39	1.70
Translator24	2088	250	11.97	1266	60.63	559	26.77	13	0.62
Translator25	3776	770	20.39	2157	57.12	842	22.30	7	0.19
Translator26	3032	199	6.56	2037	67.18	781	25.76	15	0.49
Total	**28090**	**4813**	**17.31**	**14852**	**53.44**	**8148**	**28.32**	**277**	**0.93**

Table 11: Translator-wise final decision counts along with their percentages for the cleaning task

En-Si				En-Ta			
Translator	Total Sentences	Duration (hh:mm)	Duration for 100 sentence-pairs (hh:mm)	Translator	Total Sentences	Duration (hh:mm)	Duration for 100 sentence-pairs (hh:mm)
Translator16	652	30:00	4:36	Translator27	6039	64:30	1:04
Translator17	1523	19:00	1:15	Translator28	2883	50:00	1:44
Translator18	3402	146:42	4:19	Translator29	6593	195:55	2:58
Translator19	3636	97:55	2:42	Translator30	103	12:00	11:39
Translator20	2288	28:10	1:14	Translator31	105	7:40	7:18
Translator21	2726	95:00	3:29	Translator32	151	22:00	14:34
Translator22	2669	38:35	1:27	Translator33	331	6:03	1:49
Translator23	2298	83:00	3:37	Translator34	102	6:00	5:52
Translator24	2088	92:00	4:24	Translator35	2784	39:10	1:24
Translator25	3776	125:31	3:19	Translator36	722	24:40	3:24
Translator26	3032	97:00	3:12	Translator37	199	1:30	0:45
				Translator10	1707	27:57	1:38
				Translator12	3785	61:10	1:36
				Translator38	457	9:07	1:59
				Translator39	459	6:10	1:20
				Translator40	381	6:00	1:34
Totals	**28090**	**853:18**	**33:36**	**Totals**	**26801**	**539:52**	**60:44**
Average (SD)			**3:03 (1:09)**	**Average (SD)**			**3:47 (3:57)**

Table 12: Cleaning duration analysis for Translators

	Time taken (hh:mm)	
	Translation of 100 sentence-pairs	Cleaning of 100 sentence-pairs
Translator18	03:06	4:19
Translator21	04:12	3:29
Translator26	04:25	3:12
Total Duration	**11:43**	**11:00**
Average (SD)	**03:54 (00:35)**	**03:40 (0:28)**

Table 13: Time spent to translate 100 En sentences from scratch and for cleaning of 100 `En-Si` sentence-pairs

them being Abrahamic religions (Albayrak et al., 2018), they do share some information and it is reasonable for even a human evaluator to align some of these scripture by mistake. In the last `En-Si` example, it is also evident that the sentence structure arising from the use of parentheses has played a part in aligning the wrong sentences. In row number 12 of the extended version `En-Ta`, another interesting observation is that the punctuation count (specifically the quotation marks) has also been a contributor to the misalignment.

F NMT Results

As discussed in Section 8 we report the Chrf++ as our primary evaluation metric. Apart from this we also calculate the Chrf, BLEU, and spBLEU scores as well. Since HuggingFace library doesn't support spBLEU score, we are only able to report spBLEU for vanilla-transformer. Tables 15, 16, 17, 18, and 19 contain the raw results for Figures 1, 2, 3, 4, 5 and 6 respectively.

G Domain Divergence Evaluation

We calculate the Jensen Shannon Divergence (JS-div) (Lu et al., 2020) between the training datasets (NLLB original, NLLB Cleaned top 25K, NLLB Cleand Complete (27K+), SITA Top25K, and SITA Random 25K) and the test sets (SITA and FLORES). We use the code implementation used

Extended Guidelines

1. There are three possible scenarios to handle a re-write.
a)If the two sentences are meaningful but not related, you need to translate En to Si and Si to En (so two rewrites).
b) If only En is meaningful, translate that to Si (so only one rewrite).
c) If only Si is meaningful, translate that to En (so only one rewrite).
In the case of (a),there have to be two rows now, and the decision should be selected as "**Re-write**" in **both rows**.
You can either insert a new row and copy-paste the row above it entirely (Figure 1.1) OR insert a new row and copy only the Si or En sentence as needed (Figure 1.2).

Prophecy is involved, so the claim has to be understood in a straightforward way.	ආශාව පොදි බැඳගෙන යා යුතු ගමන දුරු කළ යුතු ය, යන්න බුද්ධ දේශනාවයි.	Re-write	Prophecy is involved, so the claim has to be understood in a straightforward way.	අනාවැකිය සත්‍ය වී ඇත, එබැවින් හිමිකම සරලව තේරුම් ගත යුතුය.
Prophecy is involved, so the claim has to be understood in a straightforward way.	ආශාව පොදි බැඳගෙන යා යුතු ගමන දුරු කළ යුතු ය, යන්න බුද්ධ දේශනාවයි.	Re-write	The Buddha's teaching is that the journey to be carried with desire must be avoided.	ආශාව පොදි බැඳගෙන යා යුතු ගමන දුරු කළ යුතු ය, යන්න බුද්ධ දේශනාවයි.

Figure 1.1 - Duplicating the En and Si original sentences during Rewrite

Eventually Joseph was proven to be a prophet.	පැරණි හිබ්රූ අනාගත වක්තෘන් ගේ අනාවැකි අනුව ඔහු තමා අනාගත වක්තෘ කෙනෙක් බව හඳුන්වා දුන්නේය.	Re-write	Eventually Joseph was proven to be a prophet.	අවසානයේදී ජෝසප් අනාගතවක්තෘවරයෙකු බව ඔප්පු විය.
		Re-write	He introduced himself as a prophet according to the prophecies of the ancient Hebrew prophets.	පැරණි හිබ්රූ අනාගත වක්තෘන් ගේ අනාවැකි අනුව ඔහු තමා අනාගත වක්තෘ කෙනෙක් බව හඳුන්වා දුන්නේය.

Figure 1.2 - Keeping the En and Si original sentences as empty in the added row during Rewrite

2. In situations as shown in Figure 1.3 below, it is NOT essential to include numbers (sequence numbers/citations, etc) or punctuations that are not relevant to the sentence. Eg: [11] and " can be removed
The preferred updated sentences are shown in Figure 1.4. Note that both sides are updated.

En sentence	Si sentence	Decision	En Sentence Corrected	Si Sentence Corrected
Certainly in that there are signs for people who think."[11]	"නියත වශයෙන්ම එහි (අවබෝධ ලත් ජනයාට) සංඥා (රාශියක්) ඇත.	Update	Certainly in that there are signs for people who think."[11]	නියත වශයෙන්ම කල්පනාකාරී ජනයාට එහි සංඥා රාශියක් ඇත."[11]

Figure 1.3 - Example with punctuations/numbering that are not related to the sentence

En sentence	Si sentence	Decision	En Sentence Corrected	Si Sentence Corrected
Certainly in that there are signs for people who think."[11]	"නියත වශයෙන්ම එහි (අවබෝධ ලත් ජනයාට) සංඥා (රාශියක්) ඇත.	Update	Certainly in that there are signs for people who think.	නියත වශයෙන්ම කල්පනාකාරී ජනයාට එහි සංඥා රාශියක් ඇත.

Figure 1.4 - Preferred way of handling punctuations/numbering

3. Mark as **Delete** ONLY when both En and Si sentences are meaningless, if they contain repetitive words, (eg: No no no), or if they contain very short phrases (e.g. name of a place or a person). Otherwise, as mentioned above, the sentence should be translated to the other language

4. If the En sentence is in **spoken form**, the corrected Si sentence should also be in spoken form, and vice versa.

Figure 8: Snapshot of the Extended Guidelines given for the translators conducting the web-mining corpus cleaning

en	“ What makes you think that it will be the truth, or even accurate?”
si	මහණෙනි, තුඹලා කුමකැයි සිතන්නහු ද , රූපය නිත් යය හෝ වෙයිද?
	en: Monks, what do you think, is form constant?
en	And he opens up the refrigerator, and all he sees is the bright light .
ta	கதிரவன் தான் ஒளி யைத்தருகிறான், அனைத்தையும் காண செயகிறான்.
	en: The Sun is the one who gives light and makes everything visible.
en	God is All-knowing All-aware.
si	අපගේ ශාස්තෘ වූ බුදු රජාණන් වහන්සේ සියල්ල දන්නා සේක .
	en: Our teacher the Lord Buddha is all-knowing.
en	The two sea caves are linked, water goes in the one on the left and comes out the one on the right
ta	இரண்டு மகா கடல்கள் சங்கமிக்கும் பகுதி என்பதால், கடல் கொந்தளிப்பானது, இடதும் வலதுமாய் , முன்னும் பின்னுமாய் கப்பலை அலைக் கழிக்கும்.
	en: As it is the confluence of two great oceans, sea turbulence, will toss the ship left and right, fore and back.
en	“My Lord, the fierce beasts of the two towns are coming!”
si	"මහණෙනි, මේ දෙදෙන වනාහි බාලයෝ (අඥානයෝ) වෙත්."
	en: “Monks, these two are low (ignorant).”
en	“And I brought you some water with a straw.”
ta	8 நான் உங்களுக்குத் தண்ணீரால் திருமுழுக்குக் கொடுத்தேன் ஃ அவரோ உங்களுக்குத் தூய ஆவி-யால் திருமுழுக்குக் கொடுப்பார்" எனப் பறைசாற்றினார்.
	en: 8 I have baptized you with water, and he will baptize you with the Holy Spirit.” he declared.
en	Is that evidence that he is God ?
si	මේවට කියන්නේ දෙයියෝ සාක්කි කියලද?
	en: Are these told as gods are the witnesses?
en	The die , then, is the equivalent of a cookie cutter.
ta	மரணமும் இலைமறைக் காய் போல மறைந்து விடும்.
	en: Death will also disappear like a fruit behind the leaf.
en	and fasten them into the back of the dot.
si	තවද අපි ඔවුන්ව ඝන සෙවණෙහි ඇතුළු කරන්නෙමු .
	en: And we are admitting them in the dense shade.
en	“And we do not reveal the signs except to strike fear..”
ta	எனிமும், மிகவும் நன்றி கெட்ட, பெருந்துரோகிகளைத் தவிர வேறு எவரும் நம் அத்தாட்சிகளை நிராகரிப்-பதில்லை !
	en: However, nobody rejects the evidence except the most ungrateful, traitors !
en	No horses permitted on the said [sic] property.
si	ඔවුන්ගේ අශ්වයන්ට අලෞකික බලයක් නැත ;
	en: Their horses do not have an otherworldly power;
en	“Yes,” they said . “ you are not a person whom we doubt.”
ta	" " அவர்கள் சொல்வார்கள் ஃ " நீ எங்கள் தங்க மகனல்லவா!
	en: “ ”They will say, ”Aren’t you our golden son!
en	Thou hast given him his heart’s desire, * and hast not denied him the request of his lips .
si	`` තුඹ අපේ සුරතලුන්ට වස පෙව්වා," යන අදහස ඔහුගේ සිතට නැගිණ; දිවට නොනැගිණ .
	en: “You poisoned our pets,” the idea came to his mind; but not his lips.
en	Then introduce your family one at a time.
ta	பிறகு "இதை உமது குடும்பத்தாருக்கே உண்ணக் கொடுத்து விடுவீராக!
	en:Then “Give this to your family to eat!”
en	We have such ADD in this town (there’s something in the water!).
si	අපේ පළාතෙ තියෙන එකෙක් [පේරාදෙණි මල්වත්ත] මේ විදිහෙම පාරක් තියෙනවා.
	en: There’s a similar road at the one in our area [Peradeniya botanical garden] too.
en	between them for you to practice.
ta	அவைகளை , நீங்கள் செயல்படுத்த உங்களுக்குள் பரிசுத்த அலங்காரம் மிகவும் அவசியம்.
	en: For you to implement them, you need a holy adornment within you.

Table 14: Extended set of examples of *parallel* sentences from NLLB where the translated Si or Ta sentence has a different meaning than the original En sentences. We highlighted in colour code the pairs of semantically close words that possibly contributed to the misalignment. Correct En translation of the Si or Ta sentence is given for comparison.

Dataset		SITA				FLORES			
		Chrf	Chrf++	spBLEU	BLEU	Chrf	Chrf++	spBLEU	BLEU
NLLB	Top	17.90	15.70	2.90	0.70	22.30	20.10	5.60	1.10
	Random	8.20	6.90	0.20	0.00	7.90	6.90	0.30	0.00
	Bottom	9.30	7.80	0.30	0.00	8.40	7.40	0.30	0.00
CCMatrix	Top	24.90	22.40	8.00	1.90	24.10	21.90	8.40	1.70
	Random	5.60	4.60	0.10	0.00	5.70	4.80	0.10	0.00
	Bottom	8.70	7.10	0.10	0.00	9.80	8.00	0.00	0.00
CCAligned	Top	26.80	24.40	10.30	2.70	22.80	21.10	8.40	1.80
	Random	23.70	20.90	7.80	1.20	19.80	17.80	4.20	0.80
	Bottom	6.20	5.10	0.20	0.00	5.50	4.60	0.10	0.00
Wikimatrix	Top	21.20	18.60	5.40	0.60	23.20	20.70	7.90	1.00
	Random	10.40	8.90	0.70	0.00	11.80	10.40	1.20	0.00
	Bottom	8.50	7.10	0.30	0.00	9.00	7.70	0.40	0.00

Table 15: Chrf++ scores visualized in Figure 1 as well as other scores used for evaluation.

NMT Model		SITA				FLORES			
		Chrf	Chrf++	spBLEU	BLEU	Chrf	Chrf++	spBLEU	BLEU
vanilla-transformer	Top	24.90	22.40	8.00	1.90	24.10	21.90	8.40	1.70
	Random	5.60	4.60	0.10	0.00	5.70	4.80	0.10	0.00
	Bottom	8.80	7.10	0.10	0.00	9.80	8.00	0.00	0.00
mBART	Top	41.37	37.33	–	9.95	37.36	34.24	–	9.08
	Random	31.62	28.20	–	5.54	34.32	30.88	–	5.89
	Bottom	12.12	10.24	–	0.67	16.30	14.01	–	1.13
M2M	Top	37.61	33.85	–	8.23	34.76	31.78	–	8.03
	Random	25.66	22.89	–	4.13	29.10	26.29	–	4.75
	Bottom	10.05	8.50	–	0.71	13.89	11.99	–	1.02
NLLBm	Top	47.01	42.29	–	11.96	45.69	41.81	–	12.73
	Random	45.03	40.35	–	11.16	44.10	40.05	–	11.43
	Bottom	41.89	37.36	–	8.91	42.15	38.12	–	10.19

Table 16: Chrf++ scores visualized in Figure 2 as well as other scores used for evaluation.

Dataset Size		SITA				FLORES			
		Chrf	Chrf++	spBLEU	BLEU	Chrf	Chrf++	spBLEU	BLEU
0.1 M	En-Si	31.8	28.8	13.2	4.1	29.5	27.2	12.4	3.9
0.2 M	En-Si	34.2	30.8	15.3	4.4	32.8	30.0	15.0	4.6
0.3 M	En-Si	34.2	30.8	14.1	4.3	32.2	29.5	14.1	4.4
0.4 M	En-Si	32.6	29.4	13.6	4.1	31.9	29.3	14.0	4.1
0.5 M	En-Si	32.2	29.0	13.2	3.8	31.7	29.1	13.5	4.0
0.6 M	En-Si	32.2	29.0	12.5	3.7	32.2	29.5	13.4	4.3
0.7 M	En-Si	30.9	27.9	11.7	3.7	30.9	28.5	12.8	4.1
0.8 M	En-Si	29.7	26.8	10.9	3.4	30.1	27.6	12.0	3.9
0.9 M	En-Si	28.5	25.8	10.3	3.3	28.7	26.4	11.3	3.6
1.0 M	En-Si	27.9	25.3	10.4	3.0	28.8	26.5	11.1	3.6
1.1 M	En-Si	28.6	25.8	10.0	3.1	29.3	26.9	10.9	3.5
1.2 M	En-Si	27.6	24.9	9.7	2.9	28.8	26.4	10.8	3.6
1.3 M	En-Si	26.2	23.7	8.7	2.7	27.6	25.3	9.8	3.0
1.4 M	En-Si	27.3	24.7	9.1	2.8	28.3	25.9	10.1	3.3
1.5 M	En-Si	26.4	23.8	8.6	2.6	28.2	25.9	9.9	3.3
1.6 M	En-Si	25.0	22.7	7.6	2.5	26.6	24.4	8.6	2.8

Table 17: Raw values of ChrF++ scores visualized in Figure 3 as well as other scores used for evaluation.

by (Nayak et al., 2023). The results of the JS-div can be found in Table 20.

JS-div calculation can be described as follows. It is calculated between two distributions P and Q using the formula shown in Equation 1, where M is an equally weighted sum of $M = \frac{1}{2}P + \frac{1}{2}Q$ and $KL(\cdot||\cdot)$ represents the Kullback–Leibler divergence (Kullback and Leibler, 1951).

$$JSD(P||Q) = \frac{1}{2}KL(P||M) + \frac{1}{2}KL(Q||M) \quad (1)$$

JS-div ranges from 0 to 1 with lower values indicating that the two distributions are more similar.

Dataset		SITA				FLORES			
		Chrf	Chrf++	spBLEU	BLEU	Chrf	Chrf++	spBLEU	BLEU
SITA EnSi	SITA-top25K	46.20	43.10	29.00	15.30	21.70	18.80	3.70	0.40
	SITA-Random25K	40.80	37.90	24.00	11.30	18.90	16.30	2.10	0.20
NLLB Original EnSi	NLLB-Original-Tok25K	17.90	15.70	2.90	0.70	22.30	20.10	5.60	1.10
NLLB-Cleaned EnSi	NLLB-Cleaned-Top25K	19.20	17.00	3.70	0.80	24.30	21.90	6.60	1.70
	NLLB-Cleaned-Complete(27K+)	19.10	16.80	3.40	0.70	23.70	21.40	6.80	1.40
SITA EnTa	SITA-top25K	43.80	39.00	21.50	9.10	25.00	20.90	2.00	0.20
	SITA-Random25K	40.50	35.80	17.70	7.30	22.40	18.50	1.10	0.00
NLLB Original EnTa	NLLB-Original-Tok25K	24.30	20.50	2.60	0.40	30.40	26.20	6.30	1.10
NLLB-Cleaned EnTa	NLLB-Cleaned-Top25K	25.20	21.40	2.90	0.70	31.50	27.20	6.60	1.10
	NLLB-Cleaned-Complete(26K+)	26.40	22.50	3.30	0.70	32.70	28.40	7.50	1.20

Table 18: Chrf++ scores visualized in Figure 4 and Figure 5 as well as other scores used for evaluation.

NMT Model		SITA				FLORES			
		Chrf	Chrf++	spBLEU	BLEU	Chrf	Chrf++	spBLEU	BLEU
LASER-3	Top	24.90	22.40	8.00	1.90	24.10	21.90	8.40	1.70
	Random	5.60	4.60	0.10	0.00	5.70	4.80	0.10	0.00
	Bottom	8.80	7.10	0.10	0.00	9.80	8.00	0.00	0.00
LaBSE	Top	9.70	8.80	0.90	0.20	10.50	9.60	0.90	0.00
	Random	7.80	6.60	0.20	0.00	8.30	7.20	0.30	0.00
	Bottom	7.60	6.10	0.00	0.00	8.70	6.90	0.10	0.00

Table 19: Chrf++ scores visualized in Figure 6 as well as other scores used for evaluations.

Datasets	NLLB Top 25K BC	WikiMatrix Top 25K	CCAligned Top 25K	CCMatrix Top 25K	SITA Top 25K	NLLB Top 25K AC
SITA Test Set	0.71	0.55	0.64	0.59	0.16	0.69
FLORES Test Set	0.51	0.44	0.61	0.51	0.62	0.47

Table 20: Domain divergence between datasets for `En-Si`. `BC`- Before cleaning, `AC`- after cleaning

We calculate the JS-div for each of the test datasets (SITA and FLORES) against the following portions of the web-mined corpora: NLLB top 25k, WikiMatrix top 25k, CCAligned top 25k, SITA top 25k and NLLB top 25k.