Improving the Quality of Web-mined Parallel Corpora of Low-Resource Languages using Debiasing Heuristics

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

Abstract

Parallel Data Curation (PDC) techniques aim to filter out noisy parallel sentences from web-mined corpora. Ranking sentence pairs using similarity scores on sentence embeddings derived from Pre-trained Multilingual Language Models (multiPLMs) is the most common PDC technique. However, previous research has shown that the choice of the multiPLM significantly impacts the quality of the filtered parallel corpus, and the Neural Machine Translation (NMT) models trained using such data show a disparity across multiPLMs. This paper shows that this disparity is due to different multiPLMs being biased towards certain types of sentence pairs, which are treated as noise from an NMT point of view. We show that such noisy parallel sentences can be removed to a certain extent by employing a series of heuristics. The NMT models, trained using the curated corpus, lead to producing better results while minimizing the disparity across multiPLMs. We publicly release the source code and the curated datasets[1].

1 Introduction

Parallel data mined from the web at scale is often considered an alternative to human-created data in training Neural Machine Translation (NMT) models (Costa-jussà et al., 2022; Bañón et al., 2020). CCAligned (El-Kishky et al., 2020), CCMatrix (Schwenk et al., 2021) and ParaCrawl (Bañón et al., 2020) are examples of such web-mined corpora, which cover Low-Resource Languages (LRLs) as well. As summarised in Table 1, quality audits of these corpora reveal different types of noise. Consequently, training NMT models on such noisy parallel data leads to low-quality translations (Khayrallah and Koehn, 2018).

Parallel Data Curation (PDC) aims at extracting *high-quality* parallel sentence pairs from noisy web-mined corpora. The importance of PDC for LRLs has been emphasised with the introduction of PDC shared tasks (Sloto et al., 2023; Koehn et al., 2020, 2019). Initiated by the work of Chaudhary et al. (2019), recent PDC techniques follow a *scoring* and *ranking* mechanism using embeddings obtained from a Multilingual Language Model (multiPLM). During the *scoring* step, the semantic similarity is calculated between the source and target sentence embeddings for each sentence pair. Then the sentence pairs are *ranked* in descending order of the similarity *score*. Finally, a subset of the top-ranked sentence pairs is selected to train the NMT model. However, the quality of these top-ranked pairs depends on the chosen multiPLM (Ranathunga et al., 2024; Moon et al., 2023).

To investigate the impact of using different multiPLMs on the PDC task, we conduct an initial analysis. We obtain embeddings from three multiPLMs: LASER3 (Heffernan et al., 2022), XLM-R (Conneau et al., 2020), and LaBSE (Feng et al., 2022), calculate the semantic similarity between each parallel sentence pair in the CCMatrix (Schwenk et al., 2021) and CCAligned (El-Kishky et al., 2020) datasets and rank them in descending order. Then we train NMT models (Section 4.4) using the top-

Parallel Sentence Quality Category	A	B	C	D	E
Perfect translations	-	-	-	Y	Y
Near perfect translation	-	-	-	-	Y
Correct translation - Low quality	-	Y	-	Y	Y
Over/Under translation	-	Y	Y	Y	-
Misordered words	Y	Y	Y	Y	-
Spelling permutations	-	Y	-	Y	-
Untranslated Sentences	Y	Y	Y	-	Y
Short Sentences	Y	-	Y	-	Y
Mismatch numbers	-	Y	-	-	-
Machine Translated Sentences	-	-	Y	Y	-
Misaligned sentences	Y	Y	Y	Y	Y
Wrong Language	Y	Y	Y	Y	Y
Not a Language	Y	Y	Y	Y	Y

Table 1: Parallel sentence quality categories used in quality audits by Khayrallah and Koehn (2018) (**A**), Bane et al. (2022) (**B**), Herold et al. (2022) (**C**), Kreutzer et al. (2022) (**D**) and Ranathunga et al. (2024) (**E**).

[1]https://github.com/aloka-fernando/Heuristic_based_parallel_corpus_filtration

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28253–28270

ranked 100k sentence pairs from each corpus and report the ChrF++ scores. Experiments are carried out for English-Sinhala (En-Si), English-Tamil (En-Ta) and Sinhala-Tamil (Si-Ta) language-pairs. Sinhala, Tamil, and English belong to three distinct linguistic groups: Indo-Aryan, Dravidian, and Germanic (respectively). As shown in Figure 1, there is a significant difference, i.e. a ***disparity*** among these results, mainly for En-Si and En-Ta language pairs.

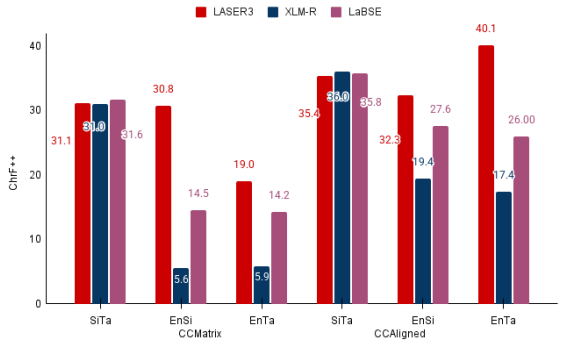


Figure 1: Baseline NMT scores using top-ranked sentence pairs from CCMatrix and CCAligned corpora, with embeddings from LASER3, XLM-R, and LaBSE.

A random inspection of the En-Si top-ranked sentences reveals that different multiPLMs prioritise different sentence characteristics when ranking parallel sentences. For example, sentence pairs ranked top with XLM-R embeddings are mostly short and have high overlapping text such as numbers, acronyms, and URLs. LaBSE embeddings also result in sentences with numbers and date overlaps, while LASER3 embeddings result in relatively better full-length sentence pairs. This observation sheds light on the disparity in NMT results shown in Figure 1 - the sentences ranked top when using XLM-R and LaBSE may have less linguistic content for NMT models to learn the translation.

In order to carry out a more systematic evaluation, we randomly select 100 sentence pairs from the top 100K parallel sentences from the aforementioned ranked corpora and carried out a human evaluation. Ranathunga et al. (2024)'s is the most comprehensive taxonomy available for this task. To better capture the noise observed during human evaluation, we extend this taxonomy (Section 3). Human evaluation results (Table 2) indicate that the mean noise percentages are 98%, 95%, and 70% when the corpora are ranked using LaBSE, XLM-R, and LASER3 embeddings, respectively. Further, the noise types in the categories of untranslated text (UN), short sentences (CS) and sentences with high-overlapping non-translatable text (CCN) could be observed, contributing to the reported percentages. Examples for these noise types are shown in Table 10 in Appendix A. We provide a detailed discussion of these results in Section 5.3. These findings suggest that inherent biases in multiPLMs lead to noisy parallel sentences being ranked highly, which in turn contributes to disparities across NMT models.

We hypothesise that some of these noisy sentences can be filtered using rule-based heuristics. Although applying heuristics is a common approach to improving the quality of parallel corpora (Sloto et al., 2023; Steingrímsson et al., 2023), the use of heuristics has not been consistent in the PDC tasks (Steingrimsson, 2023; Bala Das et al., 2023; Aulamo et al., 2020). Previous research either applied a single heuristic or a subset of commonly used heuristics during pre-processing, with threshold choices varying across studies. Further, they did not systematically analyse the impact of heuristic-based filtering on the NMT performance.

In this research, we incorporate heuristics proposed in previous studies, along with a new heuristic of our own, and conduct an empirical study to investigate whether a more refined selection of top-ranked parallel sentences can be identified by systematically combining these heuristics.

Our key contributions are as follows:

- We extend Ranathunga et al. (2024)'s parallel sentence categorization taxonomy with a new error category to capture an additional type of noise, which refers to sentence pairs with high-overlapping non-translatable text such as numbers, acronyms, URLs, etc.

- We empirically show that applying heuristics before ranking sentences based on embeddings derived from multiPLMs results in higher NMT scores, and reduces the disparity across multiPLMs.

- We conduct a systematic study to analyse the impact of rule-based heuristics in filtering noisy sentences from the web-mined corpora and identify an optimal combination of heuristics that works across corpora and languages considered in the study.

		CC	CN	CB	*C*	CS	CCN	UN	X	WL	NL	*E*	CC	CN	CB	*C*	CS	CCN	UN	X	WL	NL	*E*	CC	CN	CB	*C*	CS	CCN	UN	X	WL	NL	*E*
		Sinhala - Tamil											English - Sinhala												English - Tamil										
CCMatrix																																		
LASER3	BF	8%	27%	2%	**37%**	14%	14%	34%	1%	0%	0%	**63%**	17%	7%	4%	**28%**	7%	10%	55%	0%	0%	0%	**72%**	0%	3%	2%	**5%**	0%	0%	95%	0%	0%	0%	**95%**
	AF	16%	56%	1%	**73%**	13%	4%	10%	0%	0%	0%	**27%**	39%	39%	7%	**85%**	0%	7%	8%	0%	0%	0%	**15%**	6%	61%	20%	**87%**	0%	3%	10%	0%	0%	0%	**13%**
XLM-R	BF	1%	10%	0%	**11%**	40%	19%	29%	0%	1%	0%	**89%**	1%	0%	0%	**1%**	13%	4%	80%	2%	0%	0%	**99%**	0%	0%	2%	**2%**	3%	5%	90%	0%	0%	0%	**98%**
	AF	0%	20%	2%	**22%**	13%	29%	35%	0%	1%	0%	**78%**	3%	8%	26%	**37%**	0%	2%	53%	8%	0%	0%	**63%**	0%	39%	31%	**70%**	1%	3%	21%	4%	0%	1%	**30%**
LaBSE	BF	4%	6%	0%	**10%**	74%	7%	9%	0%	0%	0%	**90%**	13%	2%	0%	**15%**	63%	14%	8%	0%	0%	0%	**85%**	0%	9%	2%	**11%**	34%	7%	48%	0%	0%	0%	**89%**
	AF	29%	33%	0%	**62%**	2%	32%	4%	0%	0%	0%	**38%**	87%	7%	3%	**97%**	0%	1%	2%	0%	0%	0%	**3%**	36%	53%	4%	**93%**	1%	3%	2%	1%	0%	0%	**7%**
CCAligned																																		
LASER3	BF	3%	24%	3%	**30%**	34%	19%	17%	0%	0%	0%	**70%**	2%	22%	8%	**32%**	13%	30%	23%	2%	0%	0%	**68%**	2%	23%	18%	**43%**	13%	27%	17%	0%	0%	0%	**57%**
	AF	5%	79%	2%	**86%**	0%	9%	4%	1%	0%	0%	**14%**	13%	58%	14%	**85%**	0%	0%	13%	2%	0%	0%	**15%**	3%	67%	10%	**80%**	0%	8%	12%	0%	0%	0%	**20%**
XLM-R	BF	0%	0%	2%	**2%**	48%	49%	0%	0%	0%	1%	**98%**	2%	0%	0%	**2%**	72%	20%	6%	0%	0%	0%	**98%**	0%	8%	4%	**12%**	42%	16%	15%	8%	0%	7%	**88%**
	AF	20%	33%	4%	**57%**	1%	22%	19%	0%	1%	0%	**43%**	18%	18%	20%	**56%**	0%	6%	34%	4%	0%	0%	**44%**	6%	46%	30%	**82%**	0%	9%	9%	0%	0%	0%	**18%**
LaBSE	BF	0%	1%	0%	**1%**	69%	26%	3%	0%	0%	1%	**99%**	0%	1%	0%	**1%**	97%	2%	0%	0%	0%	0%	**99%**	0%	1%	0%	**1%**	97%	0%	0%	0%	0%	2%	**99%**
	AF	15%	34%	0%	**49%**	2%	43%	6%	0%	0%	0%	**51%**	45%	27%	3%	**75%**	1%	19%	5%	0%	0%	0%	**25%**	19%	45%	3%	**67%**	0%	22%	11%	0%	0%	0%	**33%**

Table 2: The Human evaluation results showing the average percentage for each annotation class for CCMatrix and CCAligned corpora for the En-Si, En-Ta and Si-Ta language pairs. The sample sentences have been obtained before (BF) and after (AF) applying the heuristics. **(*C*)** - overall correct percentage considering **CC** (perfect translation), **CN** (near perfect) and **CB** (boilerplate). **(*E*)** - overall error percentage considering **CCN** (Non-translatable overlaps), **CS** (correct but short sentence), **X** (wrong translation), **UN** (untranslated), **WL** (wrong language), **NL** (not a language).

- We conduct a human evaluation to assess the impact of noise filtering across three multiPLMs.

2 Related Work

2.1 MultiPLMs for PDC

While employing a multiPLM for PDC is common, existing research experimented with only one multiPLM at a time. For example, in the WMT2023 shared task (Sloto et al., 2023), LASER2 was utilised to set the task baseline, whereas for the same task, Steingrimsson (2023) used LaBSE. Gala et al. (2023) also used LaBSE for their work. Therefore, the disparities across multiPLMs and biases specific to each multiPLM have not come to light.

On the other hand, studies conducted by Ranathunga et al. (2024) and Moon et al. (2023) reveal that using different multiPLMs for scoring and ranking parallel corpora, and training NMT models with the top-ranked corpora, results in a disparity. Moon et al. (2023) observe that this is due to biases in multiPLMs, which tend to rank noisy parallel sentences highly. However, there has been no systematic study to identify these biases.

2.2 Identifying Noise in Web-mined Corpora

Recent research used categorical labels to annotate translation pairs, aiming at quantifying the noise types in web-mined corpora. Kreutzer et al. (2022) used their taxonomy to conduct manual audits on random samples from three web-mined datasets and reported substantial noise, specifically for LRLs. Ranathunga et al. (2024)'s taxonomy (See Table 9 in Appendix A) is an extension of Kreutzer et al. (2022)'s taxonomy. They first ranked the datasets based on embeddings from a multiPLM, and then selected random samples from the top and bottom portions and conducted a quality audit. Their human evaluation reported that the quality of the parallel sentences varies heavily depending on the selected portion. These studies primarily focus on identifying general noise types; however, their effectiveness in quantifying the specific noise types to which multiPLMs are biased has not yet been evaluated to the best of our knowledge.

2.3 Heuristic-based PDC

In existing work, the commonly used rule-based heuristics can be categorised into four groups as described below:

Deduplication-based *(Dedup)*: Removing identical duplicates from the monolingual sides is a common practice (Costa-jussà et al., 2022). Additionally deduplicating after removing non-alpha characters and punctuations (Bala Das et al., 2023) could be found as its variants. While this step has been applied during the pre-processing stage, an empirical study has not been conducted to evaluate its impact on the final NMT performance.

Length-based (*sLength*) : Gala et al. (2023) and Aulamo et al. (2023) have removed short sentences as a potential heuristic. Short sentences hinder NMT models in two ways (Koehn and Knowles, 2017): by providing insufficient syntactic and semantic information, and can result in an overfitting situation.

LID-based *(LID)*: Language Identification is used to remove fully/partially untranslated text and content in the wrong language (Steingrímsson et al., 2023; Gala et al., 2023; Zhang et al., 2020).

Ratio-based : Ratio-based heuristics identify and remove sentences that show significant structural imbalances between the source and target sentences. It is based on the assumption that, well-

aligned sentence pairs tend to maintain consistent ratios in terms of character count, word count, or token distribution. We observe three common types of ratio-based heuristics: (1) source-to-target sentence length ratio (*STRatio*) (Rossenbach et al., 2018; Gale and Church, 1993), (2) alpha-only words to sentence words ratio (*sentWRatio*) (Velayuthan et al., 2024; Aulamo et al., 2020) and (3) alpha-only character ratio with respect to the sentence characters (*sentCRatio*) (Hangya and Fraser, 2018).

However, the impact of these heuristics in isolation and as a combination had not been evaluated systematically in the context of NMT.

Ranathunga et al. (2024)	Improved Taxonomy	Revision
-	CCN	Perfect/near perfect translation where more than 30% of the overlapping content is non-translatable such as numbers/acronyms/URLs/email etc.
Short Sentences (Max 3 words)	CS	Less than 5 words on either side
Wrong Language	WL	Specifically set a threshold as 30%

Table 3: A comparison of the improved taxonomy against Ranathunga et al. (2024)'s (only showing the changes). See Table 9 in Appendix A for the full taxonomy.

3 Methodology

Improved Taxonomy: Although translation pairs can have overlapping URLs, acronyms, etc, excessive inclusions of such content in a sentence (e.g. consider the sentence *'Contact: Diane Anderson 076-8268914, info@sandnasbadenscamping.se'*[2]) do not provide meaningful content for an NMT system to learn from. However, under Ranathunga et al. (2024)'s taxonomy, such sentence pairs would likely be categorised as perfect translation-pairs **(CC)**. Therefore, we define a new noise category **CCN** (high-overlapping non-translatable text) to capture such sentence pairs.

Secondly, we consider the upper limit for short sentences as five words[3]. Finally, we improve the definition of *WL* (wrong language) to consider a threshold in determining whether a sentence pair should be marked as wrong language. Table 3 shows these changes. The complete list of these noise categories and example parallel sentences are available in Table 9 and Table 10 (respectively) in Appendix B.

Selection of Heuristics: Table 4 shows how the heuristics discussed in Section 2.3 may help in removing different noise categories. Note that a deduplication-based heuristic cannot be associated with any in the taxonomy, as it does not apply to individual sentence pairs. In addition to the deduplication strategies discussed in Section 2.3, we introduce an n-gram-based deduplication, meaning that sentences would be removed if they overlap in a consecutive n-gram text span.

Noise Category	Short Label	Rule-based Heuristic
Not a language	NL	LID, sentWRatio, sentCRatio
Wrong language	WL	LID
Untranslated	UN	LID
Short Sentences	CS	sLength
High-overlapping non-translatable text	CCN	LID, sentWRatio, sentCRatio
Wrong translation	X	STRatio (With a length difference)
Boilerplate translation	CB	STRatio

Table 4: Mapping between the noise category vs the noise mitigating heuristic.

Human Evaluation We conduct a human evaluation to quantify the noise before and after applying the heuristics. The annotator selection criteria, resources and training provided, the payment details, etc are described in Appendix A.

For each language-pair, we obtain top 1000 samples in each of the ranked corpora using embeddings obtained from LASER3, XLM-R, and LaBSE, we randomly select 100 parallel sentences for each language pair. We ask the translators to annotate each sentence pair using the taxonomy discussed in Section 3.

Each sentence pair is annotated by three translators to reduce any potential bias inherent in the individual translators. For the three annotators, the Fleiss Kappa scores are 0.833 for EnSi, 0.651 for EnTa, and 0.649 for SiTa. We note that the results for EnTa and SiTa are very close.

4 Experiments

4.1 Data

We use the language pairs, En-Si, En-Ta, and Si-Ta in our experiments. Sinhala and Tamil are morphologically rich, low-resource and mid-resourced languages (Joshi et al., 2020; Ranathunga and de Silva, 2022), respectively. Languages are selected considering the availability of human evaluators. We select CCMatrix and CCAligned as the web-mined corpora. Both these corpora include parallel data for the language pairs considered in the research. More details of these datasets and language pairs are shown in Appendix D.

For the NMT experiments, we use the *dev* and *devtest* subsets from the Flores-200 (Costa-jussà

[2]More examples are in Table 11 in Appendix B.

[3]We considered thresholds 3,4 and 5 and empirically selected this threshold as it gave the highest NMT gains. Results in Table 12 in Appendix C.

et al., 2022) dataset[4] as validation and evaluation sets, respectively. Dataset statistics are provided in Table 5.

Language-pair	CCMatrix	CCAligned	dev	devtest
En-Si	6,270,801	619,711	997	1,012
En-Ta	7,291,119	880,547	997	1,012
Si-Ta	215,966	260,118	997	1,012

Table 5: Corpus statistics.

4.2 Selection of muliPLMs

We select LASER3, XLM-R, and LaBSE for obtaining embeddings for the sentences to determine the semantic similarity. XLM-R, different to others, was trained purely on monolingual data, but has proven to be useful for cross-lingual tasks as well (Choi et al., 2021; Conneau et al., 2020). All three models include En, Si, and Ta. Details of the multiPLMs are shown in Appendix E.

4.3 Heuristic-based PDC Experiments

Each heuristic is applied independently to the source (S), target (T), and both sides (ST) of the corpora. In line with the original sentence alignment conducted for CCMatrix and CCAligned, we treat En as the source side. For Si-Ta, Si is considered the source because it is more common for a Si sentence to be translated to Ta (Farhath et al., 2018). Finally, for each multiPLM, the retained sentences are ranked in descending order with cosine similarity.

Deduplication: We consider different granularities of deduplication. i.e. identical deduplication *(dedup)*, deduplication after removing numbers only (*nums*) and removing both numbers and punctuations (*punctsNums*). Subsequently, we deduplicate considering different n-gram spans, i.e. 4-grams, 5-grams, 6-grams and 7-grams.

Length-based: We filter short sentences less than five words[3]. While some research has suggested removing extremely long sentences (Minh-Cong et al., 2023; Gala et al., 2023), we find that the percentage of longer sentences is lower and that removing them has a negligible effect. Thus, this result is not reported.

LID-based: We use a public LID model[5] (Costa-jussà et al., 2022) to predict the language of each sentence. The predicted label is then used as a standalone heuristic (*LID*) and in combination with its associated prediction probability (*LIDThresh*), with threshold of 0.7[6].

Ratio-based: For *STRatio*, 0.79-1.39, 0.87-1.62 and 0.85-1.57 were selected as thresholds for En-Si, En-Ta and Si-Ta, respectively. These thresholds are determined by calculating the mean and the standard deviation obtained for the validation set in a human-crafted trilingual dataset (Fernando et al., 2020; Ranathunga et al., 2018). Following observations of Hangya and Fraser (2018), 0.6 is selected as the threshold for *sentWRatio* and *sentCRatio*.

4.4 NMT Experiments

First, a Sentencepiece[7] tokenizer with a vocabulary size of 25000 is trained. Then we use the fairseq toolkit (Ott et al., 2019) to model and train transformer-based Seq-to-Seq NMT models until convergence. Hyperparameters used in the NMT experiments are shown in Table 13 in Appendix F. The baseline NMT models are trained on the top 100,000 sentence pairs from the ranked corpus. We use ChrF++ (Popović, 2017) to report NMT results.

5 Results and Analysis

We report the results obtained for the NMT models trained in the forward direction. The results of the experiments are in Table 6. Heuristic-wise best result is summarised in Table 14 in Appendix G. As evident from these tables, as well as from Figure 1, NMT results across different multiPLMs show a great disparity in the baseline NMT scores for En-Si and En-Ta language pairs. In the following subsections, we discuss how the use of heuristics is useful in mitigating this disparity and improving overall NMT results.

5.1 Impact of Heuristics on NMT Results

5.1.1 Impact of Deduplication-based PDC

We observe that deduplication, irrespective of the heuristic-applied side (S/T/ST), outperforms the baseline in 89% of the experiments. Overall, *dedup* considering both source and target seems to be the most effective — it outperforms the baseline in 94% of the experiments. In comparison, *dedup* target-only and source-only outperform the baseline in 89% and 83% of the experiments, respectively.

We apply our newly introduced n-gram-based deduplication *dedup+ngram*, on top of *dedup*.

[4]https://github.com/openlanguagedata/flores
[5]https://github.com/facebookresearch/fairseq/tree/nllb
[6]Thresholds below 0.7 reduce NMT results.
[7]https://github.com/google/sentencepiece

Heuristic	Applicable Side	Sinhala-Tamil						English-Sinhala						English-Tamil					
		CCMatrix			CCAligned			CCMatrix			CCAligned			CCMatrix			CCAligned		
		LASER3	XLM-R	LaBSE	LASER3	XLM-R	LaBSE	LASER3	XLM-R	LaBSE	LASER3	XLM-R	LaBSE	LASER3	XLM-R	LaBSE	LASER3	XLM-R	LaBSE
Baseline		31.08	30.99	31.63	35.36	35.97	35.79	30.76	5.55	14.49	32.33	19.39	27.57	19.02	5.86	14.20	40.13	17.40	26.00
DD	S	32.05	31.50	32.07	36.40	36.01	34.98	29.72	6.35	14.69	33.26	21.04	28.22	19.67	4.93	14.96	40.87	19.47	26.26
	T	31.39	31.44	31.73	36.26	35.86	35.96	33.81	12.59	25.97	33.66	21.41	28.32	19.48	6.87	17.96	40.13	17.90	27.79
	ST	32.26	31.10	32.25	36.41	36.08	35.32	34.01	13.80	26.18	33.47	22.22	29.49	20.32	6.45	17.53	40.56	19.83	30.01
DD-4gram	S	30.37	30.65	30.53	35.74	35.24	34.55	28.69	8.56	13.05	31.56	23.53	28.25	19.72	7.06	19.56	35.54	25.64	26.49
	T	31.00	29.90	29.39	36.05	35.98	35.44	31.79	13.60	23.66	32.86	24.95	29.05	19.82	7.08	20.23	39.83	27.44	31.18
	ST	30.86	31.13	30.80	35.28	35.36	34.64	28.72	15.17	20.45	28.15	15.45	21.37	18.15	7.00	21.37	35.02	25.70	27.41
DD-5gram	S	30.89	30.90	31.25	35.64	35.81	35.87	28.73	7.14	13.51	33.44	23.98	28.79	18.06	4.70	17.16	40.39	24.07	29.07
	T	31.24	31.55	32.10	36.26	35.87	35.23	33.98	14.01	26.23	34.10	22.27	31.10	20.15	6.75	18.78	41.12	24.05	30.26
	ST	30.78	31.53	31.35	35.64	35.94	35.44	31.95	13.87	23.07	31.60	17.10	23.52	19.61	6.25	20.12	21.77	25.22	29.36
DD-6gram	S	31.89	30.82	31.76	36.31	36.11	35.88	31.10	7.62	13.41	33.53	21.47	28.51	20.32	5.47	15.59	40.48	21.75	27.64
	T	32.51	30.41	32.29	36.35	36.23	36.01	34.21	13.98	24.91	34.24	23.63	30.23	**21.75**	6.69	20.32	40.44	20.31	30.48
	ST	31.89	30.82	31.76	35.84	35.95	35.54	33.63	14.96	24.72	33.29	15.54	25.55	20.38	7.18	20.19	41.73	24.89	31.06
DD-7gram	S	31.48	31.27	32.03	36.26	35.67	35.50	30.93	5.91	15.94	33.27	19.90	29.58	21.54	5.71	16.49	40.63	20.01	28.91
	T	31.56	31.06	30.85	36.44	36.10	35.16	34.27	13.72	25.58	32.97	22.14	28.22	20.91	7.37	**21.96**	40.49	19.18	28.69
	ST	31.48	31.27	32.03	35.74	35.90	34.82	33.93	14.95	24.95	33.63	14.58	24.96	17.56	5.98	20.71	40.94	22.16	29.40
DD+N	S	31.51	31.37	31.99	36.61	36.66	35.99	30.54	5.92	15.12	34.77	28.07	31.81	17.00	5.60	13.41	41.40	28.65	35.22
	T	31.17	30.51	32.09	36.30	36.45	36.32	33.83	14.44	25.86	34.47	27.27	31.90	17.54	6.09	19.01	41.36	28.40	35.12
	ST	31.71	31.22	31.66	36.49	36.37	36.10	33.83	14.15	26.12	34.24	28.45	31.64	19.19	5.15	18.92	41.46	30.49	35.42
DD+PN	S	31.90	31.47	31.02	36.50	36.00	36.12	30.55	6.28	16.67	34.72	27.25	31.89	18.15	5.79	15.66	41.78	30.55	35.78
	T	31.90	32.05	30.89	36.63	36.47	**36.86**	33.89	14.81	**26.31**	35.06	27.69	32.01	21.57	**8.24**	20.41	41.64	29.35	35.32
	ST	32.05	31.31	32.53	35.96	**36.71**	36.23	33.37	14.15	26.08	34.08	27.80	32.59	20.99	5.82	18.83	41.80	30.69	35.91
DD+PN+4gram	ST+T		**NA**			**NA**			**NA**		30.64	29.48	30.19		**NA**		41.82	35.90	**37.08**
DD+PN+5gram	ST + T	**32.98**	**32.73**	**32.60**	36.24	36.21	36.35	**34.50**	**16.09**	25.78	33.81	**30.33**	**32.74**		**NA**			**NA**	
DD+PN+6gram	ST + T	30.41	31.38	31.42	**36.73**	36.62	36.37		**NA**		**35.24**	28.21	31.26	19.49	6.67	20.60	**41.90**	**35.97**	35.94
DD+PN+7gram	T +T		**NA**			**NA**			**NA**			**NA**		19.57	7.55	20.89			
SL	**S**	**31.41**	**31.52**	**32.30**	36.42	36.37	36.52	32.49	6.58	20.70	33.86	26.53	32.97	17.50	5.11	18.74	41.40	27.60	36.77
	T	31.38	30.56	31.97	36.30	**36.71**	36.58	31.88	7.83	28.51	**34.88**	29.42	33.14	18.52	**6.33**	**21.73**	**41.54**	30.16	37.61
	ST	31.21	31.32	31.37	**36.47**	35.99	**36.60**	**32.82**	**8.24**	**29.96**	34.83	**29.55**	**33.50**	**19.45**	5.33	20.79	41.14	**32.67**	**38.08**
LID	**S**	**31.48**	**31.36**	**31.78**	36.05	36.03	35.64	31.00	6.23	14.69	34.39	27.33	31.73	18.44	6.93	13.43	41.80	31.41	33.95
	T	30.78	31.14	31.53	35.68	36.07	35.85	32.48	12.22	16.04	33.70	24.38	30.48	**29.59**	14.70	24.24	41.51	24.24	30.69
	ST	31.43	30.66	31.40	36.17	36.12	35.18	31.99	13.32	**16.20**	34.11	28.87	32.26	**29.59**	13.54	23.45	41.42	32.33	36.13
LT	S	30.05	31.25	31.06	35.60	35.25	34.29	30.32	7.12	15.26	**35.73**	30.86	32.69	18.98	6.02	13.06	41.60	35.25	36.29
	T	31.28	30.40	30.68	35.03	30.01	32.01	32.82	12.94	15.81	35.22	27.46	30.40	**29.59**	**15.24**	24.51	41.03	30.01	34.01
	ST	30.33	30.46	30.71	**36.73**	**36.73**	**36.80**	**32.84**	**14.08**	13.71	35.11	**32.97**	**32.88**	28.93	15.16	**25.33**	**42.63**	**38.01**	**37.40**
STRatio	-	31.74	22.80	31.34	36.39	35.74	35.30	31.09	5.20	15.40	33.47	24.05	30.21	**20.52**	5.40	**18.29**	40.91	22.71	28.61
sentWRatio	S	30.65	30.62	32.03	36.17	35.77	35.54	**31.50**	**7.40**	10.86	**34.15**	25.97	**31.35**	19.42	5.79	13.93	**42.05**	29.70	35.53
	T	30.71	31.59	31.34	36.24	36.17	**36.46**	30.99	6.39	15.13	33.51	26.93	30.47	18.61	5.65	11.08	41.87	30.06	35.54
	ST	31.93	31.56	30.98	**36.44**	**36.72**	36.01	30.64	7.00	**15.50**	33.85	**28.73**	31.17	18.99	4.82	14.08	41.05	**30.88**	**35.77**
sentCRatio	S	31.67	31.24	31.14	35.94	36.18	35.86	30.15	7.05	14.46	34.06	21.52	30.10	17.47	6.22	13.83	40.68	22.48	29.37
	T	30.98	31.21	31.93	36.36	35.43	35.85	30.65	5.83	15.28	33.64	23.14	29.05	19.90	**6.78**	12.51	40.78	19.63	29.42
	ST	**32.28**	**31.90**	**32.04**	36.33	35.60	36.11	30.85	6.45	14.64	33.60	23.84	29.70	19.54	6.45	10.79	41.76	21.82	30.82
Combined Heuristics																			
DD+PN+ngram (SiTa-CCMatrix n=5, SiTa-CCAligned n= 7 EnSi-CCMatrix/CCAligned n=5, EnTa-CCMatrix n=7, EnTa-CCAligned n=6)																			
+sLength	T + ST	30.17	29.02	29.99	36.32	36.81	36.61	35.03	21.70	26.32	35.68	33.49	34.43	30.29	19.44	29.85	42.84	39.36	40.16
+LT	T + ST	31.49	30.13	30.68	36.58	36.37	37.02	35.42	19.58	32.43	34.77	32.58	34.72	20.53	7.52	23.35	42.68	38.45	39.60
+sentWRatio	T+S	31.37	30.55	30.92	**36.83**	36.75	36.30	33.99	15.76	24.92	33.97	31.40	32.72	21.67	8.23	24.58	42.11	37.47	38.07
+SL+LT	T + ST	29.28	30.85	29.96	36.47	36.81	36.88	35.70	23.92	32.77	34.97	34.92	35.60	30.65	20.86	31.49	42.85	41.17	41.31
+SL+sentWRatio	T + ST + ST	31.45	**32.65**	31.17	36.60	**36.85**	36.32	35.71	18.93	32.53	35.45	33.42	33.82	22.46	9.11	23.82	41.97	40.07	40.06
+SL+LT+sentWRatio	T+ST+ST+S	29.81	29.53	29.73	**36.83**	36.66	**37.03**	**36.10**	23.84	**33.94**	36.15	34.50	**35.67**		**NA**		**43.47**	**41.74**	41.06
+SL+LT+sentWRatio>0.8	T+ST+ST+ST	28.70	28.39	28.34	36.20	36.60	35.89	35.66	**24.18**	33.19	**36.26**	**35.66**	35.42		**NA**		42.08	40.56	**42.02**
+SL+LT+sentCRatio	T+ST+ST+ST	**32.64**	31.30	**32.28**		**NA**			**NA**			**NA**			**NA**			**NA**	
+SL+LT+STRatio	T+ST+ST+STR		**NA**			**NA**			**NA**			**NA**		**30.67**	**23.36**	**31.80**		**NA**	

Table 6: NMT results obtained after applying heuristics in isolation and in combination in the ablation study. The values in bold indicate the highest NMT score obtained for a given heuristic class or from the heuristic combination. The values underlined are the highest among the individual heuristics. Highlighted in green are the overall best values. Here **DD+PN** is *deduplication+punctNums*, **SL** is *sLength* and **LT** is *LIDThresh*. **NA** would be when the particular experiment is not applicable for that language pair or the dataset.

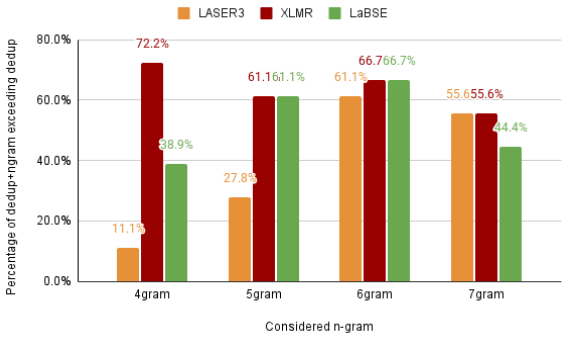


Figure 2: Percentage of *dedup+ngram* experiments exceeding the best result of *dedup* for each *multiPLM*

We find that for each result column, there is a *dedup+ngram* result that outperforms the best results obtained with the corresponding *dedup* result. To observe the impact of the *n* value on the NMT result, we plot Figure 2 showing the percentage of *n-gram* experiments exceeding the highest *dedup* result, with respect to the multiPLM (language-wise result is in Figure 4 of the Appendix G). We observe a consistent pattern - n=5 or 6 perform the best in a majority of cases. We believe that 4-gram results in an overly aggressive deduplication. However, the exact n-gram depends on the corpus characteristics.

We observe that *dedup+punctNums* outperforms *dedup+nums* and *dedup* in 78% and 89% of the experiments (respectively), proving that *(dedup+punctNums)* to be more impactful. Finally, we analyse the impact of

dedup+puntsNums+ngram[8]. Compared to other *dedup* combinations[9], *dedup+punctNums+ngram* produce the best result across 67% of the experiments.

5.1.2 Impact of Length-Based PDC

sLength surpasses the baseline in 89% of the experiments. Therefore, we can conclude that *sLength* is favourable as a heuristic. When analysing the side on which the heuristic is applied, we observe that applying it to both ST is the most effective (similar to *dedup*), followed by T and then S (56%, 28%, and 17% of the experiments, respectively).

Recall that our manual inspection noticed that **XLM-R and LaBSE tend to prioritise shorter sentences**. This observation is affirmed by the *sLength* results. For En-Si and En-Ta, this heuristic resulted in substantial gains for many of the XLM-R and LaBSE experiments, while it shows marginal improvements for LASER3.

5.1.3 Impact of LID-Based PDC

With LID-based PDC, we observe substantial improvements for XLM-R and LaBSE, except for Si-Ta, for which the gains are marginal. Gains are reported by LASER3 as well, though not very significant. The highest gain of +20.61 ChrF++ is reported for XLM-R for the CCAligned-EnTa corpus, while a gain of +11.40 is reported by LaBSE for the same corpus.

NMT models trained after applying *LID* outperform the baseline in 85% of the experiments, while *LIDThresh* outperforms *LID* in 72% of the experiments. Therefore, we conclude that *LIDThresh* would be the most suitable heuristic. We observe that the gains for *LIDThresh* with respect to *LID* are least with Si-Ta with 50% while for En-Si and En-Ta it is 83% each. We assume that this is due to a limitation with the LID model, which is not optimised for Si and Ta.

5.1.4 Impact of Ratio-based PDC

We observe that *STRatio*, *sentWRatio* and *sentCRatio* produce NMT gains over baseline for 56%, 69% and 80% of experiments, respectively. Among the three ratio-based heuristics, *sentWRatio* outperforms both *STRatio* and *sentCRatio* in 67% of the experiments, making it the most effective for noise reduction. In contrast, *STRatio* and *sentCRatio* exceed the performance of the other two heuristics in only 11% and 22% of the cases, respectively. Maximum gains are reported for the EnTa-CCAligned corpus, with +1.92, +13.48, and +9.77 ChrF++ for LASER3, XLM-R, and LaBSE respectively.

5.2 Summary of Heuristic-based PDC

1. Impact of the Individual Heuristics on the NMT Results: The LID-based heuristic emerged as the most impactful in 44% of the experiments. Deduplication and sentence-length heuristics were the most effective in 33% and 17% of the experiments, respectively. The ratio-based heuristic alone did not hold superior results compared to others, making it the least impactful single heuristic. There is no single heuristic that consistently produces the best gains for all scenarios. For LASER3, XLM-R, and LaBSE, we observe the highest ChrF++ gains of +10.57, +20.61, and +11.40 for CCMatrix-EnTa and CCAligned-EnTa and CCAligned-EnTa corpora, respectively.

2. Impact of Combined Heuristics: The combination of heuristics produced the best score, compared to the best-performing individual heuristics, except for the CCMatrix-SiTa. In this, the combination reduced the dataset size by 54% (Table 15 in Appendix G), resulting in only 98k parallel sentences. We suspect this reduction in dataset size and the residual noise could result in the reduced score.

CCMatrix-SiTa gains across multiPLMs are marginal. The same pattern holds with LASER3 across En-Ta and En-Si languages, with the exception of CCMatrix-EnTa corpus. Since LASER3 is already trained with OPUS data (Tiedemann and Thottingal, 2020)[10], it is believed that LASER3 is already optimised (Moon et al., 2023) towards ranking the CCMatrix and CCAligned corpora better. This is further affirmed with the relatively higher NMT scores of the baseline experiments.

For the rest of the ranked corpora, the best NMT scores were reported when the combined heuristic was applied. It was noted the combination always had heuristics, ***dedup+punctNums+(n)gram+sLength+LIDThresh*** as a common combination for 77% of the experiments. The specific ***n***-value was dependent on the dataset. In the combination producing the best gains, the ratio-based heuristic was different based on the dataset/language pair. In 61% of the cases, the best scores were obtained with ***sentWRatio***,

[8] *puntsNums* and *ngram* has been applied on top of *dedup*.
[9] *dedup*, *dedup+ngram*, *nums*, and *dedup+punctNums*
[10] https://opus.nlpl.eu/

while ***STRatio*** yielded the best results in 16% of the experiments.

The combination performed best without the LID heuristic only in CCAligned-SiTa with XLM-R. However, we include *LIDThresh* into the combination for two reasons. (1) Even with *LIDThreshold*, the score only lags by (-0.19) ChrF++ scores compared to the best NMT score, which is negligible. (2) The most influential individual heuristic is LIDThresh-based for most cases.

Therefore we recommend the heuristic combination, ***dedup+punctNums+(n)gram+sLength +LIDThresh+sentWRatio*** to be applied as the rule-based heuristic combination for the PDC task.

3. Reducing the Disparity Across multiPLMs: We calculate the disparity (Δ) as the difference in NMT scores between LASER3 and the XLM-R or LaBSE scores. Equation 1 shows the baseline disparity calculation, where $BL_{multiPLM}$ refers to the baseline NMT score obtained using embeddings from either XLM-R or LABSE during ranking.

$$\Delta_{Baseline} = BL_{LASER3} - BL_{multiPLM} \quad (1)$$

The disparity for each heuristic is calculated with respect to LASER3 and the best NMT score obtained from either XLM-R or LaBSE. The results are shown in Table 7. Additionally, we calculate the disparity reduction percentage ($\Delta\ Reduction(\%)$) as defined in Equation 2. Here, $\Delta_{heuristic}$ is the disparity corresponding to the best-performing NMT models after applying the respective heuristic or the optimal combination.

$$\Delta\ Reduction(\%) = \frac{\Delta_{\text{baseline}} - \Delta_{\text{heuristic}}}{\Delta_{\text{baseline}}} \times 100\% \quad (2)$$

In Table 7[11], we observe that the disparity has been reduced on average by 95.87% for CCAligned-XLM-R/LaBSE and CCMatrix-LaBSE, compared to LASER3. The disparity between CCMatrix-XLM-R EnSi/EnTa models, compared to LASER3 was reduced by only 48%, meaning that XLM-R ranked corpus contains noise that cannot be mitigated by the heuristics alone. However, our hypothesis holds in most cases, and it is safe to say that heuristic-based PDC mitigates the bias brought in by the multiPLM.

[11] Since the disparity observed was marginal for the SiTa pair, we exclude this language pair from our discussion.

Heuristic	LASER3 vs XLM-R		LASER3 vs LaBSE	
	Δ (ChrF++)	Δ Reduction (%)	Δ (ChrF++)	Δ Reduction (%)
CCMatrix				
English - Sinhala				
Baseline	25.21		16.27	
Deduplication - based	18.41	26.97%	8.72	46.40%
Sentence Length - based	24.58	2.50%	2.86	82.42%
LID -based	18.76	25.59%	16.64	-2.27%
Ratio-based	24.10	4.40%	16.00	1.66%
Combined Heuristics	11.92	**52.72%**	2.16	**86.72%**
English - Tamil				
Disparity	13.16		4.82	
Reduction in disparity (dedup)	13.33	-1.29%	-0.39	108.09%
Reduction in disparity (sLength)	13.12	0.30%	-2.28	147.30%
LID	14.35	-9.04%	4.26	11.62%
Ratio-based	13.74	-4.41%	2.23	53.73%
Combined Heuristics	7.31	**44.45%**	-1.13	**123.44%**
CCAligned				
English - Sinhala				
Baseline	12.94		4.76	
Deduplication Best	4.91	62.06%	2.50	47.48%
Reduction in disparity (sLength)	5.33	58.81%	1.38	71.01%
LID	2.76	78.67%	2.85	40.13%
Ratio-based	5.42	58.11%	2.80	41.18%
Combined Heuristics	0.60	**95.36%**	0.59	**87.61%**
English - Tamil				
Disparity	22.73		14.13	
Reduction in disparity (dedup)	5.93	73.91%	4.82	65.89%
Reduction in disparity (sLength)	8.87	60.98%	3.46	75.51%
LID	4.62	79.67%	5.23	62.99%
Ratio-based	11.17	50.86%	6.28	55.56%
Combined Heuristics	1.73	**92.39%**	1.45	**89.74%**

Table 7: ***Disparity*** (Δ) in ChrF++ points, among the NMT models (XLM-R/LaBSE) with the best scores with respect to LASER3 after applying the individual/combined heuristics. The Δ ***Reduction (%)*** is this disparity as a percentage of the baseline disparity.

5.3 Human Evaluation Results

As shown in Table 2, heuristic-based PDC had reduced the noise in the top-ranked samples consistently, irrespective of the language pair and the considered multiPLM. Some gains are quite significant — for example, the amount of correct pairs (C) improvement for CCMatrix-EnSi (LaBSE), CCMatrix-EnTa (LaBSE), and CCAligned-EnSi (LaBSE) were 82%, 82%, and 74% respectively. CS, CCN and UN noise categories are noted to be contributing towards this noise percentage. However, after heuristic filtration, the error drops drastically. On average, for LASER, XLM-R, and LaBSE, the final error percentages are 2.17%, 2.50% and 1.0%, respectively. The evaluation further reveals that the residual noise are from untranslated (UN) and overlapping text (CCN) classes. These findings reveal that rule-based heuristics, are not effective in eliminating those types of noise. As a result, we would need to employ an alignment model similar to Steingrímsson et al. (2023) or Minh-Cong et al. (2023) to remove such residual noise from the corpus.

In conclusion, the human evaluation results indicate that the heuristic-based PDC approach is beneficial for parallel sentence ranking in two key ways. First, it produces the top-ranked sentence

pairs from multiPLM to be qualitatively comparable. Secondly, it removes the noisy parallel sentences that cause the disparity among the NMT systems trained using ranked corpora based on multiPLMs.

6 NMT Performance based on Training Data Size

We further investigate the effect of the training dataset size on the final NMT performance. We sampled the top 50k, 100k, 150k, and 200k from each of the ranked corpora after applying the heuristic-combination, and train NMT models. Results are in Figure 5 of the Appendix H.

For Si-Ta, moderate but consistent improvements are observed from 50K to 100K samples for both CCMatrix and CCAligned corpora, irrespective of the multiPLMs. However, after heuristic-based filtering, the resulting Si-Ta parallel corpus contained only about 100K sentence pairs (Table 15 in Appendix G). Therefore, NMT experiments with 150K and 200K sentence pairs could not be conducted due to insufficient data.

For En-Ta, NMT scores declined after 100K, while for En-Si, an improvement was observed until 200K, except for the CCMatrix ranked using XLM-R embeddings. Hence, we suspect that noise may be more prominent beyond 100K for En-Ta. In conclusion, increasing corpus size can enhance NMT performance, but only when the underlying data is sufficiently clean and well-aligned.

7 Conclusion

In this research, we empirically analysed the disparity between the NMT systems trained with the web-mined corpora ranked using embeddings derived from multiPLMs. With a human evaluation, we showed that this disparity is due to different types of noise creeping into the top-ranked portion of corpora when different multiPLMs are used.

We made use of rule-based heuristics to remove this noise. After a systematic evaluation of heuristics, we were able to identify optimal heuristic combinations that resulted in higher NMT scores. Therefore, for anyone planning to use web-mined corpora, our recommendation is to first filter out noisy sentences using heuristics and then to do ranking on the embeddings derived from the multiPLM. In contrast to the recent PDC work (Steingrimsson, 2023; Minh-Cong et al., 2023) that employs several deep learning model-based rigorous filtration pipelines, our technique is much simpler.

Human evaluation indicated that even after applying heuristics, some noise remains. In future, we plan to apply techniques such as classification-based approaches to remove such noise, or to apply translation post-editing to improve the sentence quality.

8 Limitations and Ethical Concerns

8.1 Limitations

We found that the LID was suboptimal for identifying Si or Ta languages. Therefore, such models would need to be optimised or better LID models would need to be used, such that they are effective in predicting the language label correctly. Furthermore, due to the lack of human annotators, we had to limit this analysis to only three language pairs. We can extend this study to more low-resource language pairs provided that we can find suitable annotators.

8.2 Ethical Concerns

We use publicly available datasets. Due to the dataset size, we did not have the resources or funding to check all the curated parallel sentences manually for offensive content. Fernando et al. (2020) provided the human-crafted dataset to be used in this research. Further, we do not disclose or share any personal details of the annotators publicly or with any other party. We have offered the annotators the standard rate and have settled the payments in full. The only details we disclose are in Appendix A.

Acknowledgments

This research was funded by the Google Award for Inclusion Research (AIR) 2022 received by Surangika Ranathunga and Nisansa de Silva. We would also like to thank and acknowledge the National Languages Processing Centre (NLPC), at the University of Moratuwa for providing the GPUs to execute the experiments related to the research.

References

Mikko Aulamo, Ona de Gibert, Sami Virpioja, and Jörg Tiedemann. 2023. Unsupervised feature selection for effective parallel corpus filtering. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 31–38. European Association for Machine Translation.

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A configurable parallel corpus filtering toolbox. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156. Association for Computational Linguistics.

Sudhansu Bala Das, Atharv Biradar, Tapas Kumar Mishra, and Bidyut Kr. Patra. 2023. Improving multilingual neural machine translation system for indic languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–24.

Fred Bane, Celia Soler Uguet, Wiktor Stribiżew, and Anna Zaretskaya. 2022. A comparison of data filtering methods for neural machine translation. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 313–325. Association for Machine Translation in the Americas.

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.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 261–266. Association for Computational Linguistics.

Hyunjin Choi, Judong Kim, Seongho Joe, Seungjai Min, and Youngjune Gwon. 2021. Analyzing zero-shot cross-lingual transfer in supervised nlp tasks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9608–9613. IEEE.

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. Association for Computational Linguistics.

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

Nisansa de Silva. 2025. Survey on Publicly Available Sinhala Natural Language Processing Tools and Research. *arXiv preprint arXiv:1906.02358v25*.

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. Association for Computational Linguistics.

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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891. Association for Computational Linguistics.

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*.

Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Viktor Hangya and Alexander Fraser. 2018. An unsupervised system for parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 882–887. Association for Computational Linguistics.

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. 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. 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. 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. 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. Association for Computational Linguistics.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39. 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.

Nguyen-Hoang Minh-Cong, Nguyen Van Vinh, and Nguyen Le-Minh. 2023. A fast method to filter noisy parallel data WMT2023 shared task on parallel data curation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 359–365. Association for Computational Linguistics.

Hyeonseok Moon, Chanjun Park, Seonmin Koo, Jungseob Lee, Seungjun Lee, Jaehyung Seo, Sugyeong Eo, Yoonna Jang, Hyunjoong Kim, Hyounggyu Lee, et al. 2023. Doubts on the reliability of parallel corpus filtering. *Expert Systems with Applications*, 233:120962.

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. 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. 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. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. 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. Association for Computational Linguistics.

Surangika Ranathunga, Nisansa De Silva, Velayuthan Menan, Aloka Fernando, and Charitha Rathnayake. 2024. Quality does matter: A detailed look at the quality and utility of web-mined parallel corpora. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 860–880. Association for Computational Linguistics.

Surangika Ranathunga, Fathima Farhath, Uthayasanker Thayasivam, Sanath Jayasena, and Gihan Dias. 2018. Si-ta: Machine translation of sinhala and tamil official documents. In *2018 National Information Technology Conference (NITC)*, pages 1–6. IEEE.

Nick Rossenbach, Jan Rosendahl, Yunsu Kim, Miguel Graça, Aman Gokrani, and Hermann Ney. 2018. The RWTH Aachen University filtering system for the WMT 2018 parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 946–954. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. 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. Association for Computational Linguistics.

Steve Sloto, Brian Thompson, Huda Khayrallah, Tobias Domhan, Thamme Gowda, and Philipp Koehn. 2023. Findings of the WMT 2023 shared task on parallel data curation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 95–102. Association for Computational Linguistics.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023. Filtering matters: Experiments in filtering training sets for machine translation. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 588–600. University of Tartu Library.

Steinthor Steingrimsson. 2023. A sentence alignment approach to document alignment and multi-faceted filtering for curating parallel sentence pairs from web-crawled data. In *Proceedings of the Eighth Conference on Machine Translation*, pages 366–374. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480. European Association for Machine Translation.

Menan Velayuthan, Dilith Jayakody, Nisansa De Silva, Aloka Fernando, and Surangika Ranathunga. 2024. Back to the stats: Rescuing low resource neural machine translation with statistical methods. In *Proceedings of the Ninth Conference on Machine Translation*, pages 901–907. Association for Computational Linguistics.

Yudhanjaya Wijeratne, Nisansa de Silva, and Yashothara Shanmugarasa. 2019. Natural language processing for government: Problems and potential. Technical report, LIRNEasia.

Boliang Zhang, Ajay Nagesh, and Kevin Knight. 2020. Parallel corpus filtering via pre-trained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8545–8554. Association for Computational Linguistics.

A Human Evaluation

Selection of the Annotators: We select annotators for this task who have translation or machine translation-related experience for a minimum of 2 years and are fluent in both languages of the specific language pair assigned to them. Table 8 shows the years of experience and the qualifications of those annotators who conducted this task.

Annotator	Experience (Years)	Qualification
Annotator 01	22	Diploma in Translation And Interpretation
Annotator 02	9	BSc.(Hons) in Information Technology
Annotator 03	5	MBBS
Annotator 04	4	BA (Hons) in Translation
Annotator 05	3	BSc (Hons) Engineering sp. in Computer Science and Engineering
Annotator 06	2.5	Diploma in Translation And Interpretation
Annotator 07	2.5	BSc Eng (Hons) Electrical & Electronics Engineering
Annotator 08	2.5	BSc (Hons) Engineering sp. in Electrical Engineering
Annotator 09	2	Bachelor of Industrial Information Technology

Table 8: Annotator details with the years of experience and their qualifications.

Resources Provided and Training: All the annotators have had prior experience with a similar task (Ranathunga et al., 2024). However, for this annotation work, we provide them with the definitions of the noise categories (Table 9) along with example sentence pairs (Table 10) and the guideline in terms of a flowchart (Figure 3). First, we asked them to do a sample of 30 sentences as a training on the task, and review it. Then, the 1200 sentence pairs to be annotated were shared with each annotator via Google Sheets to be completed.

Parallel Corpus Categorization Code: Description
CC: **Perfect Translation-pair** Source and target sentences are translation pairs of each other.
CN: **Near Perfect Translation-pair** Perfect translation pairs. Just a few spelling, grammar, punctuation, or unnecessary characters have to be handled.
CB: **Low-quality Translation-pair** A full sentence or phrase, but a low-quality (boilerplate) translation. Includes under/over translations.
CS: **Short Translation Content** Less than 5 words. Translation-wise, correct, but only a short phrase or a few words.
CCN: **High-overlapping non-translatable text** Perfect or near-perfect translation pair, but with overlapping content like numbers, acronyms, or URLs. Sentences longer than 5 words with high overlap.
X: **Wrong Translation** Source and target sentences are in the correct languages, but semantically unrelated. Not true translations.
UN: **Untranslated Text** The source or target is copied from its counterpart (partial or full). Overlapping untranslated content exceeds 30%. It could have been translated/transliterated.
NL: **Not a Language** At least one side is not linguistic content.
WL: **Wrong Language** Either the source or the target (or both) is not in the expected language. Up to 30% of acceptable content may be tolerated.

Table 9: Improved Ranathunga et al. (2024)'s taxonomy to categorize parallel sentences to identify biases in multiPLMs.

Compensation: They were paid the standard rate in Sri Lanka for each sentence pair they annotated or were offered co-authorship of this paper.

B Improved Noise Taxonomy

In this research, we extend the taxonomy of Ranathunga et al. (2024) by introducing the noise category **CCN** (high-overlapping non-translatable

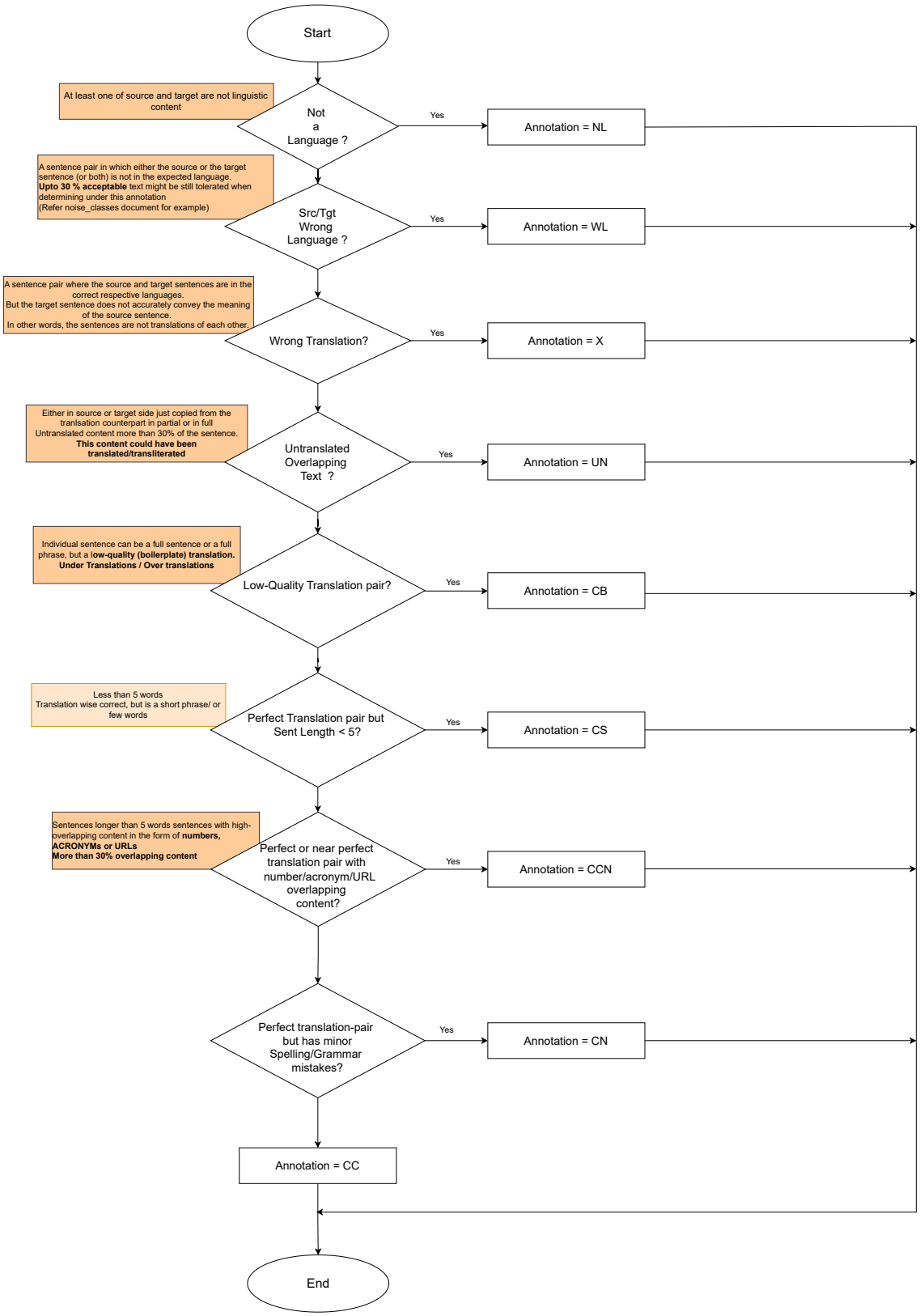


Figure 3: Shows the annotation guideline document in terms of a flow chart. This shows the priority of the noise category to be selected prior to declaring the annotation class.

Language Pair	Ranked multiPLM	Source Sentence	Target Sentence
Correct Codes			
Perfect Translation Pair (CC): Semantically equivalent sentences.			
Si-Ta	LaBSE	ජාතික නීති සම්මන්ත්‍රණය 2013 – ශ්‍රී ලංකා නීතිඥ සංගමය	தேசிய சட்ட மாநாடு 2013 – இலங்கை சட்டத்தரணிகள் சங்கம்
Low-Quality Translation Pair (CB): Weak Translation pair. Under / Over translations			
En - Ta	LASER3	Where will you be in the next five, ten or fifteen years?	கே: அடுத்த ஐந்து, பத்து அல்லது பதினைந்து ஆண்டுகளில் நீங்கள் எங்கே இருப்பீர்கள்?
Error Codes			
High-overlapping non-translatable text (CCN): Translation wise perfect or near perfect, with non-translatable text such as numbers/acronyms/URLs/emails.			
En-Ta	LaBSE	Servanet.se, info@servanet.se, or by phone 0200-120 035.	Servanet.se, info@servanet.se, அல்லது தொலைபேசி மூலம் 0200-120 035.
Incorrect Translation (X) : Both languages correct. But has translation errors			
En - Si	LaBSE	Ample storage room and slots for credit cards, IDs and Cash	ක්‍රෙඩිට් කාඩ්, හැඳුනුම් පත් සහ මුදල් සඳහා ඇති තරම් ගබඩා කාමරය සහ මඳබව
Untranslated Text (UN): either in source or target side just copied from the translation counterpart			
En - Si	XLM-R	What do you mean when you say “Your comment is awaiting moderation?”	මොකෝ විචාරක තුමා මගේ කමෙන්ට් එක තාම" Your comment is awaiting moderation.
En - Ta	XLM-R	Effective Pixels: 16.0 million (Image processing may reduce the number of effective pixels.)	ஆப்டிகல் சென்சார் ரெசொலூஷன் 20.1 million (Image processing may reduce the number of effective pixels)
Not a language (NL): at least one of source and target is not linguistic content			
En - Si	LaBSE	2.0mm2 / 900mm 2.5mm2 / 900mm 4.0mm2 / 900mm	1.5mm2 / 900mm 2.0mm2 / 900mm 2.5mm2 / 900mm 4.0mm2 / 900mm
En - Ta	LaBSE	HQCCWM750GAH6A	பதிவிறக்க: HQCCWM750GAH6A.pdf
Wrong Language (WL): Source and Target side are linguistic content. However, the source, target, or both sides a not in the expected language.			
En - Si	LaBSE	Ի պատից պարոն Գոլջիի:	ගොල්ගි මහතා විසින් එතුමාගේ නමින් ම නම් කෙරිණි.
Short Sentence (CS): Correct Translation, but the number of tokens on the Source or Target side is less			
En - Si	LaBSE	Account Number	ගිණුම් අංකය
En - Si	LaBSE	11 July 2015.	11 ජූලි 2015.
En - Ta	XLM-R	July 21:	ஜூலை 21:
Si - Ta	XLM-R	ගණනය: 40 / 2, 40 / 3, 30 ආදිය.	எண்: 40 / 2, 40, 3, 30 முதலியன

Table 10: Example parallel sentences from the En-Si, En-Ta and Si-Ta, identified during human evaluation. The translation error in the language pair is highlighted in pink.

En	2 September 1948 – 8 July 1994
Si	2 සැප්තැම්බර් 1948 – 8 ජූලි 1994
En	V2.77: French Translation, finally! [August 22, 2009]
Ta	V2.77: பிரஞ்சு மொழிபெயர்ப்பு, இறுதியாக! [ஆகஸ்ட் 22, 2009]
Si	සම්බන්ධතා: ඩයැන් ඇන්ඩර්සන් 076-826 89 14, info@sandnasbadenscamping.se
Ta	தொடர்பு: டயான் ஆண்டர்ஸன் 076-826 89 14, info@sandnasbadenscamping.se

Table 11: Example parallel sentences which will be separately identified under the new noise category *CCN*

text) to capture the type of noise that multiPLMs are biased to rank highly. Table 11 shows examples for this noise type. The definitions of all the categories in our improved taxonomy are given in Table 9. Then, in Table 10, we show examples of parallel sentences which fall under each of these categories.

C Sentence Length Threshold

To determine the optimal threshold for sentence length filtering, we filter sentences with fewer than 3, 4, and 5 words on the source (S) side, target (T) side, and both sides. Similar to other PDC experiments, we then rank the parallel sentence pairs by computing embeddings using XLM-R, LABSE

and LASER. Finally, we train NMT models using the top 100k-ranked data. We conduct these experiments using the CCAligned corpus for En-Si and En-Ta language pairs. The ChrF++ results are shown in Table 12. We observe sentence length threshold of 5 yields the best results consistently. Therefore, we select this threshold in all subsequent experiments.

	Side	LASER3	XLM-R	LaBSE
English - Sinhala				
Baseline		32.33	19.39	27.57
Sentence Length (flt < 5)	S	33.86	26.53	32.97
	T	**34.88**	29.42	33.14
	ST	34.83	**29.55**	**33.50**
Sentence Length (flt < 4)	S	34.25	23.20	31.07
	T	33.99	23.06	31.30
	ST	34.69	25.64	31.55
Sentence Length (flt < 3)	S	34.50	25.94	31.91
	T	34.14	26.55	32.49
	ST	34.06	27.16	32.76
English - Sinhala				
Baseline		40.13	17.40	26.00
Sentence Length (flt < 5)	S	41.40	27.60	36.77
	T	**41.54**	30.16	37.61
	ST	41.14	**32.67**	**38.08**
Sentence Length (flt < 4)	S	40.85	24.28	36.03
	T	41.15	26.28	37.05
	ST	41.52	30.40	37.75
Sentence Length (flt < 3)	S	40.90	22.13	33.56
	T	41.42	24.37	36.22
	ST	40.58	25.45	36.60

Table 12: NMT results in ChrF++ for different sentence length thresholds.

D Selection of Languages and Datasets

This section contains details on the selected languages and the web-mined corpora considered under the study.

Sinhala is an Indo-Aryan language spoken primarily in Sri Lanka by the Sinhalese majority. It exhibits complex morphological structures, including rich inflectional and derivational processes, but is classified as a low-resource language due to the scarcity of linguistic resources and tools (de Silva, 2025; Ranathunga and de Silva, 2022).

Tamil, a Dravidian language with a rich literary history, is spoken by Tamil communities in Sri Lanka, India, and the global diaspora. Unlike Sinhala, Tamil benefits from a relatively larger digital presence, but it still faces challenges in NLP applications due to morphological complexity, agglutinative grammar, and resource limitations in certain domains (Wijeratne et al., 2019).

We obtain the publicly released CCMatrix and CCAligned datasets from the OPUS collection (Tiedemann and Thottingal, 2020)[12]. Both these datasets support the language pairs, En-Si, En-Ta and Si-Ta, which we consider in this research.

CCMatrix (Schwenk et al., 2021) is a web-mined parallel corpus extracted using LASER2-based sentence embeddings to align bitext. While it provides large-scale data, it is highly noisy due to the *global mining* approach to determine alignments, resulting in misaligned or low-quality translations.

CCAligned (El-Kishky et al., 2020) extracts bitext from Common Crawl[13] using document-level and sentence-level alignment based on multilingual embeddings. Though it improves alignment quality over global bitext-mined corpora, it still contains significant noise, requiring careful filtering for reliable use.

E Selection of multiPLMs

We include the details on the three multiPLMs considered in this study.

LASER3 (Heffernan et al., 2022) (L=12, H=1024, A=4, P=250M)[14] is a multiPLM favourable for bitext mining and cross-lingual tasks. It improves over previous LASER2 versions by supporting more languages and enhancing alignment quality, but still faces challenges in low-resource settings.

XLM-R (Conneau et al., 2020) (L=12, H=768, A=6, P=278M) is a transformer-based multiPLM trained on massive amounts of text using masked language modelling. It achieves strong cross-lingual performance but struggles with low-resource languages due to limited training data.

LaBSE (Feng et al., 2022) (L=12, H=768, A=12, P=471M) is a BERT-based model optimised for multilingual sentence embeddings and bitext retrieval. It provides high-quality cross-lingual representations and is favourable for cross-lingual tasks.

F NMT Experiments

The experiments are conducted on a NVIDIA Quadro RTX6000 GPU with 24GB VRAM. The hyperparameters used during training, along with the training parameters, are shown in Table 13. We conduct training on the NMT systems for 100

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

[13] https://commoncrawl.org/

[14] No. of Layers, Hidden Layer Dimensions, No of Attention Heads, and Total number of parameters are defined by L, H, A, and P respectively.

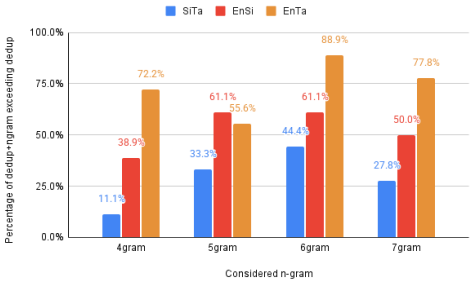


Figure 4: Percentage of *dedup+ngram* experiments exceeding the best result of *dedup* with respect to the Language-pair.

epochs with early stopping criteria and report the results using ChrF++. ChrF++ was chosen over the conventional multi-BLEU (Papineni et al., 2002) and sacreBLEU (Post, 2018) because character-level evaluation is more suitable for the considered languages, Sinhala and Tamil, which are morphologically rich in nature.

Hyperparameter	**Argument value**
encoder/decoder Layers	6
encoder/decoder attention heads	4
encoder-embed-dim	512
decoder-embed-dim	512
encoder-ffn-embed-dim	2048
decoder-ffn-embed-dim	2048
dropout	0.4
attention-dropout	0.2
optimizer	adam
Adam β_1, Adam β_2	0.9, 0.99
warmup-updates	4000
warmup-init-lr	1e-7
learning rate	1e-3
batch-size	32
patience	6
fp16	True

Table 13: Training parameters for NMT experiments.

G Results Analysis

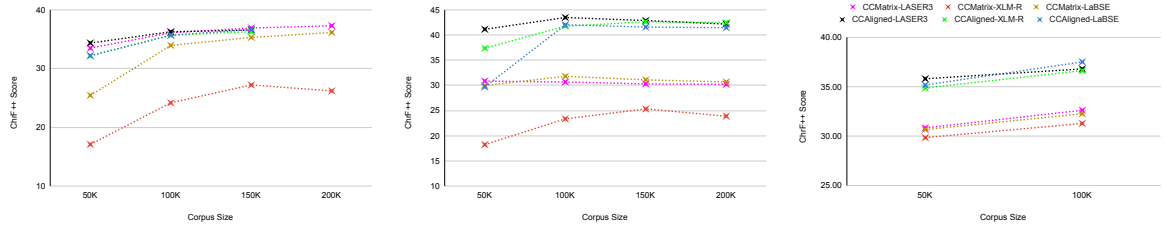
We show the individual and the combined heuristic which produced the best NMT gains with respect to each multiPLM, dataset and the language pair in the Table 14. The Figure 4 shows the percentage of *dedup+n-gram* experiments exceeding the highest *dedup*, with respect to the language-pair. Finally, we show the final dataset sizes after applying the heuristic along with the percentage reduction in Table 15.

H NMT Performance on Training Dataset

Figure 5 shows the variation in NMT results by varying the dataset size with 50k, 100, 150k and 200k for each language-pair.

Heustic/s	CCMatrix			Heustic/s	CCAligned		
	LASER3	XLM-R	LaBSE		LASER3	XLM-R	LaBSE
Sinhala →Tamil							
Baseline	31.08	30.99	31.63	Baseline	35.36	35.97	35.79
Deduplication-based Heuristics							
DD+PN+5gram (ST+T)	**32.98**	**32.73**	**32.60**	DD+puntsNums (T)	36.63	36.47	**36.86**
				DD+PN (ST)	35.96	**36.71**	36.23
				DD+PN+7gram	**36.73**	36.62	36.37
Length-based Heurics							
(S)	31.41	31.52	32.30	(T)	36.30	**36.71**	36.58
				(ST)	**36.47**	35.99	**36.60**
LID-based Heuristics							
LID (S)	31.48	31.36	31.78	LIDThresh (ST)	**36.73**	**36.73**	36.80
Ratio-based Heuristics							
sentCRatio (ST)	32.28	31.90	32.04	sentWRatio (T)	36.24	36.17	**36.46**
				sentWRatio (ST)	**36.44**	**36.72**	36.01
Combined Heuristics							
DD+PN+5gram++sentCharRatio	31.45	**32.65**	31.17	DD+PN+7gram++sentWRatio	36.60	**36.85**	36.32
DD+PN+5gram++LIDThresh+sentCharRatio	**32.64**	31.30	**32.28**	DD+PN+7gram++LIDThresh+sentWRatio0.8	**36.83**	36.66	**37.03**
Gain (DD-heuristic vs Baseline)	1.90	1.74	0.97		1.37	0.50	1.07
Gain (sLength vs Baseline)	0.33	0.53	0.67		0.94	0.74	0.79
Gain (LID-heuristic vs Baseline)	0.40	0.37	0.15		1.37	0.76	1.01
Gain (Raio heuristic vs Baseline)	1.20	0.91	0.41		1.08	0.75	0.67
Gain (Overall individual Heuristic vs Baseline)	**1.90**	**1.74**	**0.97**		**1.37**	**0.76**	**1.07**
Gain (Combined Heuristic vs Baseline)	**1.56**	**1.66**	**0.65**		**1.47**	**0.88**	**1.24**
Gain (Combined vs Individual)	**-0.34**	**-0.08**	**-0.32**		**0.10**	**0.12**	**0.17**
English →Sinhala							
Baseline	30.76	5.55	14.49	Baseline	32.33	19.39	27.57
Deduplication-based Heuristics							
DD+puntsNums (T)	33.89	14.81	**26.31**	DD+puntsNums+5gram (ST+T)	33.81	**30.33**	**32.74**
DD+puntsNums+5gram (T+T)	**34.50**	**16.09**	25.78	DD+puntsNums+6gram (ST+T)	**35.24**	28.21	31.26
Length-based Heurics							
sLength (ST)	32.82	8.24	**29.96**	sLength (T)	**34.88**	29.42	33.14
				sLength (ST)	34.83	**29.55**	**33.50**
LID-based Heuristics							
LID	31.99	13.32	**16.20**	LIDThresh (S)	**35.73**	30.86	32.69
LIDThresh	**32.84**	**14.08**	13.71	LIDThresh (ST)	35.11	**32.97**	**32.88**
Ratio-based Heuristics							
sentWRatio (S)	**31.50**	**7.40**	10.86	sentWRatio (S)	**34.15**	25.97	**31.35**
sentWRatio (ST)	30.64	7.00	**15.50**	sentWRatio (ST)	33.85	**28.73**	31.17
Combined Heuristics							
DD+PN+5gram +sLength +LIDThresh +sentWRatio	**36.10**	23.84	**33.94**	DD+PN+5gram+sLength+LIDThresh+sentWRatio	36.15	34.50	**35.67**
DD+PN+5gram+sLength+LIDThresh+sentWRatio>0.8	35.66	**24.18**	33.19	DD+PN+5gram+sLength+LIDThresh+sentWRatio>0.8	**36.26**	**35.66**	35.42
Gain (Dedup-heuristic vs Baseline)	3.74	10.54	11.82		2.91	10.94	5.17
Gain (sLength vs Baseline)	2.06	2.69	15.47		2.55	10.16	5.93
Gain (LID-heuristic vs Baseline)	2.08	8.53	1.71		3.40	13.58	5.31
Gain (Raio heuristic vs Baseline)	0.74	1.85	1.01		1.82	9.34	3.78
Gain (Overall individual Heuristic vs Baseline)	**3.74**	**10.54**	**15.47**		**3.40**	**13.58**	**5.93**
Gain (Combined Heuristic vs Baseline)	**5.34**	**18.63**	**19.45**		**3.93**	**16.27**	**8.10**
Gain (Combined vs Individual)	**1.60**	**8.09**	**3.98**		**0.53**	**2.69**	**2.17**
English →Tamil							
Baseline	19.02	5.86	14.20	Baseline	40.13	17.40	26.00
Deduplication-based Heuristics							
DD-7gram (T)	18.15	7.37	**21.96**	DD+puntsNums+4gram (ST+T)	41.82	35.90	**37.08**
DD+puntNums (T)	**21.57**	**8.24**	20.41	DD+puntsNums+6gram (ST+ST)	**41.90**	**35.97**	35.94
Length-based Heurics							
sLength (T)	18.52	**6.33**	**21.73**	sLength (T)	**41.54**	30.16	37.61
sLength (ST)	**19.45**	5.33	20.79	sLength (ST)	41.14	**32.67**	**38.08**
LID-based Heuristics							
LIDThresh (T)	**29.59**	**15.24**	24.51	LIDThresh ST)	**42.63**	**38.01**	**37.40**
LIDThresh (ST)	28.93	15.16	**25.33**				
Ratio-based Heuristics							
STRatio	**20.52**	5.40	**18.29**	sentWRatio (S)	**42.05**	29.70	35.53
sentCRatio (T)	19.90	**6.78**	12.51	sentWRatio (ST)	41.05	**30.88**	**35.77**
Combined Heuristics							
DD+PN+7gram+slength+LIDThresh+STRatio	**30.67**	**23.36**	**31.80**	DD+PN+6gram+sLength+LIDThresh+sentWRatio	**43.47**	**41.74**	41.06
				DD+PN+6gram+sLength+LIDThresh+sentWRatio > 0.8	42.08	40.56	**42.02**
Gain (Dedup-heuristic vs Baseline)	2.55	2.38	7.76		1.77	18.57	11.08
Gain (sLength vs Baseline)	0.43	0.47	7.53		1.41	15.27	12.08
Gain (LID-heuristic vs Baseline)	10.57	0.92	10.31		2.50	20.61	11.40
Gain (Raio heuristic vs Baseline)	1.50	0.92	4.09		1.92	13.48	9.77
Gain (Overall individual Heuristic vs Baseline)	**10.57**	**2.38**	**10.31**		**2.50**	**20.61**	**12.08**
Gain (Combined Heuristic vs Baseline)	**11.65**	**17.50**	**17.60**		**1.92**	**12.30**	**16.02**
Gain (Combined vs Individual)	**1.08**	**8.12**	**6.47**		**0.84**	**3.73**	**3.94**

Table 14: Shows the **best** NMT performance produced using the individual heuristics as well as combinations of heuristics. The values in bold indicate the highest NMT score obtained for a given heuristic or heuristic combination. The values underlined are the highest among the individual heuristics. Highlighted in green are the overall best values. Here **DD+PN** is *Deduplication+punctNums*, **SL** is *sLength* and **LT** is *LIDThresh*.



(a) En-Si (b) En-Ta (c) Si-Ta

Figure 5: ChrF++ scores of NMT systems trained by varying the training dataset size.

Heuristic	Applicable Side	Sinhala - Tamil				English - Sinhala				English - Tamil			
		CCMatrix		CCAligned		CCMatrix		CCAligned		CCMatrix		CCAligned	
		Dataset Size	% Reduction	Dataset Size	% Reduction	Dataset Size	% Reduction	Dataset Size	% Reduction	Dataset Size	% Reduction	Dataset Size	% Reduction
Baseline		215965		260119		6270800		619730		7291118		880568	
DD	S	189654		250038	3.88%	6146819	1.98%	570768	7.90%	6378607	12.52%	797071	9.48%
	T	209461		247176	4.98%	3242950	48.28%	562088	9.30%	4754106	34.80%	780355	11.38%
	ST	183904		243384	6.43%	3176145	49.35%	537581	13.26%	4060447	44.31%	736212	16.39%
DD-4gram	S	176590	18.23%	189218	27.26%	4171884	33.47%	403993	34.81%	4802654	34.13%	440248	50.00%
	T	196538	9.00%	207603	20.19%	2751819	56.12%	423752	31.62%	4271550	41.41%	549621	37.58%
	ST	172440	20.15%	184159	29.20%	2035282	67.54%	355790	42.59%		100.00%	365648	58.48%
DD-5gram	S	188457	12.74%	217733	16.29%	4486108	28.46%	481021	22.38%	5104643	29.99%	558504	36.57%
	T	204045	5.52%	226738	12.83%	3071693	51.02%	499307	19.43%	4516819	38.05%	638600	27.48%
	ST	185915	13.91%	216389	16.81%	2374578	62.13%	446838	27.90%	3319964	54.47%	487465	44.64%
DD-6gram	S	196194	9.15%	232604	10.58%	5383674	14.15%	528525	14.72%	5309083	27.18%	639961	27.32%
	T	200310	7.25%	237467	8.71%	3142124	49.89%	539513	12.94%	4590987	37.03%	681186	22.64%
	ST	196194	9.15%	233792	10.12%	2859756	54.40%	505429	18.44%	3489629	52.14%	570403	35.22%
DD-7gram	S	200899	6.98%	240704	7.46%	5701801	9.07%	554025	10.60%	5718913	21.56%	679750	22.81%
	T	204485	5.32%	244007	6.19%	3170538	49.44%	561285	9.43%	4631486	36.48%	703950	20.06%
	ST	200898	6.98%	246104	5.39%	3021457	51.82%	538898	13.04%	3745771	48.63%	611821	30.52%
DD+N	S	182386	15.55%	260119	0.00%	6105433	2.64%	505828	18.38%	6337285	13.08%	683882	22.34%
	T	201551	6.67%	218980	15.82%	3225979	48.56%	502971	18.84%	4722644	35.23%	675627	23.27%
	ST	176040	18.49%	216238	16.87%	3158067	49.64%	476379	23.13%	4031459	44.71%	631485	28.29%
DD+PN	S	180380	16.48%	215100	17.31%	5931349	5.41%	494778	20.16%	6194331	15.04%	668849	24.04%
	T	198352	8.16%	212341	18.37%	3197186	49.01%	492801	20.48%	4666158	36.00%	660902	24.95%
	ST	173804	19.52%	207197	20.35%	3130297	50.08%	465617	24.87%	3987832	45.31%	616702	29.97%
DD+PN+4gram	ST+T							289248	53.33%				
DD+PN+5gram	ST + T	167022	22.66%	187250	28.01%	3044520	51.45%	380146	38.66%				
DD+PN+6gram	ST + T	169784	21.38%	189060	27.32%			428939	30.79%	4547759	37.63%	464424	47.26%
DD+PN+7gram	T +T			196221	24.56%					4620008	36.64%		
SL	S	150094.00	30.50%	188061	27.70%	5088747	18.85%	411474.00	33.60%	6498956	10.86%	595057	32.42%
	T	100799.00	53.33%	161363	37.97%	3670963	41.46%	377708.00	39.05%	4267495	41.47%	517516	41.23%
	ST	96264.00	55.43%	157978	39.27%	3341564	46.71%	348829.00	43.71%	4134919	43.29%	491207	44.22%
LID	S	192377.00	10.92%	241617	7.11%	6200355	1.12%	479589.00	22.61%	7210848	1.10%	669260	24.00%
	T	186720.00	13.54%	241863	7.02%	6066681	3.26%	575298.00	7.17%	6800923	6.72%	794143	9.81%
	ST	178276.00	17.45%	231619	10.96%	6010065	4.16%	457639.00	26.16%	6743988	7.50%	625281	28.99%
LT	S	181470.00	15.97%	227791	12.43%	6120792	2.39%	398272.00	35.73%	6120793	16.05%	564870	35.85%
	T	172726.00	20.02%	222290	14.54%	5990169	4.48%	546472.00	11.82%	5990170	17.84%	731484	16.93%
	ST	162777.00	24.63%	208644	19.79%	5877142	6.28%	377579.00	39.07%	5877143	19.39%	518010	41.17%
STRatio	-	170168.00	21.21%	229101	11.92%	4293239	31.54%	459473.00	25.86%	4051888	44.43%	679820	22.80%
sentWRatio	S	199788	7.49%	246908	5.08%	6232528	0.61%	546460	11.82%	7231531	0.82%	743624	15.55%
	T	196812	8.87%	250151	3.83%	6198124	1.16%	552798	10.80%	7176111	1.58%	745221	15.37%
	ST	193989	10.18%	245161	5.75%	6177212	1.49%	531963	14.16%	7138854	2.09%	717159	18.56%
sentCRatio	S	212287	1.70%	224031	13.87%	6262297	0.14%	594169	4.12%	7252444	0.53%	832611	5.45%
	T	212877	1.43%	218726	15.91%	6261991	0.14%	596346	3.77%	7281050	0.14%	851343	3.32%
	ST	211661	1.99%	215151	17.29%	6257079	0.22%	588310	5.07%	7247298	0.60%	826866	6.10%
Combined Heuristics													
DD+PN+ngram (SiTa-CCMatrix n=5, SiTa-CCAligned n= 7 EnSi-CCMatrix/CCAligned n=5, EnTa-CCMatrix n=7, EnTa-CCAligned n=6)													
+SL	T+ST	117198	45.73%	143919	44.67%	2245307	64.19%	240086	61.26%	3462458	52.51%	321003	63.55%
+LT	T+ST	130831	39.42%	162543	37.51%	2880530	54.06%	239144	61.41%	2876818	60.54%	306352	65.21%
+sentWRatio	T+S	154028	28.68%	170715	34.37%	2993730	52.26%	337634	45.52%	3377988	53.67%	427587	51.44%
+SL+LT	T+ ST	99207	54.06%	127284	51.07%	2200195	64.91%	180731	70.84%	4330140	40.61%	241679	72.55%
+SL+sentWRatio	T+ST+ST	116344	46.13%	127035	51.16%	2241513	64.25%	224542	63.77%	2794637	61.67%	298141	66.14%
+SL+LT+sentWRatio	T+ST+ST+S	127188	41.11%	117970	54.65%	2197629	64.95%	177711	71.32%	2726087	62.61%	237105	73.07%
+SL+LT+sentWRatio>0.8	T+ST+ST+ST	179984	16.66%	99311	61.82%	2149037	65.73%	161869	73.88%	2203591	69.78%	214639	75.62%
+SL+LT+sentCRatio	T+ST+ST+ST	98894	54.21%	NA		NA		NA		NA		NA	
+SL+LT+STRatio	T+ST+ST+STR	82866	61.63%	NA		NA		NA		2129744	70.79%	NA	

Table 15: Shows the final corpus sizes after applying heuristics along with the reduction percentage. Here **DD+PN** is *Deduplication+punctNums*, **SL** is *sLength* and **LT** is *LIDThresh*. **NA** corresponds to the experiments that are not applicable for the language-pair.