Knowledge and Information Systems (2023) 65:571–612
https://doi.org/10.1007/s10115-022-01761-x

REGULAR PAPER

Exploiting bilingual lexicons to improve multilingual embedding-based document and sentence alignment for low-resource languages

Aloka Fernando[1] · Surangika Ranathunga[1] · Dilan Sachintha[1] · Lakmali Piyarathna[1] · Charith Rajitha[1]

Received: 27 April 2022 / Revised: 4 July 2022 / Accepted: 12 September 2022 /
Published online: 17 October 2022

Abstract
Neural machine translation systems trained on low-resource languages produce sub-optimal results due to the scarcity of large parallel datasets. To alleviate this problem, parallel corpora can be mined from the web. Two key tasks in a parallel corpus mining pipeline are web document alignment and sentence alignment. Effective approaches for these tasks obtained vector representations of the documents (or sentences) belonging to the two languages and determine the alignment between the documents (or sentences) based on a semantic similarity scoring mechanism. Recently, document or sentence representations obtained from pre-trained multilingual language models (PMLMs) such as LASER, XLM-R and LaBSE have significantly improved the benchmark scores in diverse natural language processing tasks. In this study, we carry out an empirical analysis of the effectiveness of these PMLMs of the document and sentence alignment tasks in the context of the low-resource language pairs Sinhala–English, Tamil–English and Sinhala–Tamil. Further, we introduce a weighting mechanism based on small-scale bilingual lexicons to improve the semantic similarity measurement between sentences and documents. Our results show that both document and sentence alignment can be further improved using our weighting mechanism. We have also compiled a gold-standard evaluation benchmark dataset for document alignment and sentence alignment tasks for the

Aloka Fernando, Dilan Sachintha, Lakmali Piyarathna and Charith Rajitha have contributed equally to this work.

✉ Aloka Fernando
alokaf@cse.mrt.ac.lk

Surangika Ranathunga
surangika@cse.mrt.ac.lk

Dilan Sachintha
dilansachintha.16@cse.mrt.ac.lk

Lakmali Piyarathna
lakmali.16@cse.mrt.ac.lk

Charith Rajitha
rajitha.16@cse.mrt.ac.lk

[1] Department of Computer Science and Engineering, University of Moratuwa, Katubedda, Sri Lanka

considered language pairs. This dataset (https://github.com/kdissa/comparable-corpus) and the source code (https://github.com/nlpcuom/parallel_corpus_mining) are publicly released.

Keywords Document alignment · Sentence alignment · Low-resource languages · Neural machine translation · Parallel corpus mining

1 Introduction

Neural machine translation (NMT) systems trained on the transformer architecture [1] produce state-of-the-art results when large parallel datasets are available. However, in low-resource settings, the same architectures produce sub-optimal results [2].

Mining for parallel corpora from the web is one commonly explored solution to alleviate the parallel data scarcity problem [3]. Wikipedia, news websites and official government/institution websites are sources that are likely to contain translations of each other, to be considered for parallel corpus mining. However, for low-resource languages, the data on the web are noisy and not of good quality, which results in a noisy parallel data when automatically mined [4]. Therefore, it is essential to implement parallel corpus mining techniques that produce quality parallel data for low-resource languages.

Document alignment and sentence alignment are important tasks in the parallel corpus mining pipeline [5]. Document alignment refers to the process of identifying web documents that contain translations of each other, which are known as *comparable corpora* [6]. Early work on document alignment mainly relied on feature-based techniques that exploited URL meta data [7–9], HTML document structure [10] or machine translation-based techniques [11, 12]. However, these were outperformed by techniques that used vector representations of documents. Recent research in this line exploited document representations derived from Pre-trained Multilingual Language Models (PMLMs), which proved to be far superior to previous techniques [13].

The objective of sentence alignment is to find parallel sentences in the already identified comparable corpora or aligned documents. Existing techniques for sentence alignment were based on sentence-level features [14], information retrieval-based techniques [12], using supervised classifiers [15] and using machine translation [16]. However, more successful techniques were based on multilingual sentence embeddings [17, 18].

Currently available PMLMs to derive multilingual sentence embeddings include LASER [19], XLM-R [20], mBERT [21] and LaBSE [18]. However, except for Rajitha et al. [22] who compared LASER and XLM-R embeddings for document alignment, and Feng et al. [18] who compared LaBSE and LASER for sentence alignment, to the best of our knowledge, there has been no comprehensive evaluation of the effectiveness of these embeddings for document or sentence alignment tasks for low-resource languages. Moreover, these models are known to provide sub-optimal results for languages that are under-represented in these PMLMs [23]. These languages turn out to be those that have already been classified as low-resource languages [24]. Thus, it is important to investigate and identify ways to improve the performance of these PMLMs for document and sentence alignment in the context of low-resource languages.

In this paper, we exploit the use of bilingual lexicons to improve the semantic similarity measurement of the sentence embeddings derived from PMLMs for the tasks of document and sentence alignment. Note that bilingual lexicons can be considered as parallel data that are in the form of short phrases.

Our document alignment system is based on the work of El-Kishky and Guzmán [13]. They derived sentence embeddings using LASER and calculated the semantic distance between documents in source and target languages using the Cross-lingual Sentence Movers Distance algorithm. Our sentence alignment system is based on the work of Artetxe and Schwenk [25]. They first obtained sentence embeddings of all source and target side sentences using LASER and calculated margin-based cosine similarity over nearest neighbours.

For both these techniques, we introduce a new weighting mechanism to improve the semantic distance measurement, by utilizing existing bilingual lexicons. Our bilingual lexicons include a bilingual dictionary, glossary, designation list and person name lists. Additionally, we exploit the effectiveness of our technique considering XLM-R [20] and LaBSE [18] in addition to LASER [19] multilingual embeddings. Thus, our work also serves as the first comparative study of the performance of these three multilingual models for these tasks.

We experiment with three language pairs: Sinhala–Tamil, Sinhala–English and Tamil–English. We have compiled a gold-standard human-annotated benchmark evaluation set for document alignment and sentence alignment tasks, in these three language pairs. The considered languages belong to three distinct language families (English (En)—Indo European, Tamil (Ta)—Dravidian and Sinhala (Si)—Indo Aryan), and Sinhala and Tamil are morphologically rich low-resource languages. Thus, this dataset is a much tougher benchmark compared to other multilingual datasets [26] that only focused on a pair of high-resource related languages. We publicly release this dataset [1] in the hope that it would serve in further research in this domain. This is the first manually curated dataset for the considered three languages.

Our experiments show that the use of bilingual lexicons improves the performance of the selected document and sentence alignment techniques, with the largest gains in the context of the LASER sentence representations.
Thus, the contributions of this work are as follows:

1. We introduce a weighting scheme based on bilingual lexicons to improve the semantic similarity measurement of the document and sentence representations derived from pre-trained multilingual models, for the document and sentence alignment tasks, respectively.
2. We conduct an empirical evaluation of the performance of sentence representations derived from LASER, XLM-R and LaBSE [2], for document and sentence alignment tasks in the context of low-resource languages.
3. We publicly release the gold-standard human-annotated benchmark evaluation datasets for the document and sentence alignment tasks in the context of three low-resource language pairs: English–Sinhala, English–Tamil and Sinhala–Tamil.

The rest of the work is organized as follows. Related work is covered in Sect. 2. In Sect. 3, we explain our approach for creating the benchmark evaluation sets and declare the bilingual lexicons used. Our lexicon-based solution is presented in Sect. 4. Results are reported in Sect. 5, with a further analysis of the results in Sect. 6. Finally, the conclusion and future work are included in Sect. 7.

[1] https://github.com/kdissa/comparable-corpus.

[2] mBERT was not considered since it does not include Sinhala.

2 Related work

A typical parallel corpus mining pipeline follows a sequence of tasks, namely: crawling of website data, alignment of web documents, sentence alignment and parallel sentence filtration [5]. In our study, we focus on document alignment and sentence alignment tasks.

2.1 Document alignment

Automatic document alignment refers to determining the likelihood that two documents are translations of each other. Early work on document alignment was mostly based on metadata, such as URL-based properties [7, 8], publication date [9] and HTML document structure/tags [10, 27, 28]. These were further extended with topic modelling techniques [29]. Although meta-data is a strong indication of document alignment, this alone is not effective, as the alignment properties mainly lie in the textual content. Further, such properties cannot be generalized across different domains and web sources.

In translation-based document alignment methods, the objective is to identify a strong signal that a document in the source language is the translation of another in the target language. Some of these techniques incorporated a bilingual dictionary and checked the existence of bilingual lexical terms in the documents [30]. Some others considered the alignment information at word level [31–33], phrase level [12, 34] or considered the existence of the n-best translated terms [11, 35] in the documents. Few techniques translated the non-English document to English and determined the alignment based on an MT evaluation metric [29, 36, 37]. Even though translation-based methods were able to score well in document alignment, their performance highly depends on the accuracy of the alignment algorithm or the translation system used. Thus, in low-resource settings, these may produce sub-optimal results.

Vector representation-based techniques first derive a vector representation for the documents in the two languages and employ a semantic distance measurement metric to determine the semantic similarity between the documents. Document pairs that obtain a semantic similarity value above a pre-determined threshold are considered to be comparable. Bag-of-words, TF-IDF [38–41] and word n-grams [42] were among the early solutions to derive document representations.

Very recently, El-Kishky and Guzmán [13] used PMLMs to derive a vector representation of each of the documents in the considered languages. Then, the distance between these document vectors was calculated to determine the aligned document pair. They used the LASER pre-trained embeddings [19] to derive document embeddings. El-Kishky and Guzmán [13] experimented for high-resource, mid-resource and also low-resource languages. As mentioned earlier, this is the baseline for our research, and more information on this technique can be found in Sect. 4.1.1.

2.2 Sentence alignment

Sentence alignment refers to the process of identifying parallel sentence pairs that are partial or complete translations of each other. Early work on parallel sentence alignment was based on sentence-length ratio [43, 44], which was purely statistical. However, when the correlation between source and target languages decreases, the performance of this approach drops rapidly [45]. Subsequent techniques considered bilingual dictionaries [14], word/phrase alignment probabilities between the sentences [12, 32] and phrase/sentence alignments coupled with bilingual suffix trees [46]. Stefanescu et al. [47] addressed sentence alignment as an

information retrieval problem, while Munteanu and Marcu [15] trained a supervised classifier to determine the alignment. Some other techniques were based on machine translation [16, 48], where the non-English source sentences were translated to English and IR techniques were used to identify the aligned candidate sentence [49–51].

Recent work for sentence alignment was based on sentence representations by means of word embeddings or sentence embeddings. Here, first, the sentence embeddings were obtained for source and target sentences. Then, using a semantic similarity measurement, the aligned sentence pairs were identified. Initial work employed word-based embeddings trained on bi-directional Recurrent Neural Networks (RNNs) [52], Deep Averaging Networks (DANs) [53], bi-gram driven network architectures [54] and auto-encoders [55]. Hybrid techniques were also adopted, where a supervised classifier was used on top of the embedding-based semantic similarity calculation to determine the alignment [55, 56].

To optimize the results of the sentence alignment task, either the sentence representations should be enhanced or the semantic similarity distance scoring should be improved. Following the former path, Artetxe and Schwenk [19] used LASER supervised multilingual embeddings, while Kvapilíková et al. [17] experimented with XLM-based unsupervised multilingual embeddings. The choice of the sentence similarity measurement technique has been largely unsupervised (cosine similarity was the simplest one employed). However, this simple method is sub-optimal, and improved semantic similarity measurements had also been proposed [25, 54, 57]. We use Artetxe and Schwenk [25] as the baseline for the sentence alignment system, and this similarity measurement technique is further discussed in Sect. 4.2.1.

The final step is parallel sentence filtration, with the objective of removing any noisy parallel sentence pairs that had crept into the mined parallel corpus due to the noise in web data itself or due to the limitations in the preceding steps. However, this step is not explored in the scope of the study.

2.3 Pre-trained multilingual language models (PMLMs)

As discussed in the previous two sections, sentence representations derived from PMLMs have been vital in the success of recent document and sentence alignment techniques. Artetxe and Schwenk [19] used parallel data to train a shared encoder (available via the LASER toolkit), which had performed well on massive-scale parallel corpus extraction projects such as ParaCrawl [5], wikiMatrix [58] and ccMatix [59].

Current state-of-the-art multilingual models had been trained on the Transformer architecture [1]. Commonly used mBERT [21] (104 languages) and XLM-R [20] (100 languages) models had been trained on monolingual data with Masked Language Modelling (MLM) objective. Yang et al. [60] trained multilingual embeddings on parallel data using a bi-directional dual encoder with an additive margin softmax objective. The latter had been used in the work of LaBSE [18] (109 languages) which had been trained using both monolingual and parallel data with the MLM and Translation Language Modelling (TML) objectives, producing the state-of-the-art results for sentence alignment. However, LaBSE had not been evaluated in the context of low-resource languages for the tasks of document and sentence alignment.

2.4 Evaluating document alignment and sentence alignment

Datasets to evaluate document alignment and sentence alignment techniques have been introduced by several shared tasks. For example, Buck and Koehn [6] provided a hand-aligned dataset for evaluation of the document alignment task. However, this dataset was limited only to English and French. Rather than creating manually aligned datasets for the task of sentence alignment on comparable corpora, Zweigenbaum et al. [61] artificially injected parallel sentences into the comparable corpus. Their dataset also focused only on four language pairs, Chinese–English, French–English, German–English, and Russian–English. Some shared tasks did not present manually aligned datasets [26, 62]. Rather, the performance was evaluated by using the identified parallel sentences on a downstream NMT task.

3 Dataset

In this section, we describe the approach taken to create the gold standard evaluation set for the document alignment and sentence alignment tasks (Sect. 3.1). Afterward, we describe the human evaluation conducted to evaluate the quality of our gold-standard evaluation datasets (Sect. 3.2). In Sect. 3.3, we outline the bilingual lexicons used in this research.

3.1 Preparing document and sentence alignment evaluation datasets

Our research focuses on Sinhala, Tamil and English languages, which are the official languages of Sri Lanka. We selected the news websites that publish content in all these three languages as comparable web sources. The selected web sites were Hiru News [3], NewsFirst[4], Army News [5] and ITN [6]. We considered data from 2013 January up to April 2021.
During pre-processing, news content in paragraphs of each web page was merged into a single string, and the text contained in the image and video tags were discarded. Further, we have removed very short news documents that contained tokens less than fifty.

Army, Hiru and Newsfirst websites publish news in all three languages with the same content coverage, document structure, order of sentences and information ow. Hence, for most of the English documents, the exact translations were available in Sinhala and Tamil documents. However, for ITN News, this was different. We observed that the English article was not always available for the corresponding news in Sinhala and Tamil language articles, and for the ones with translations, there was a low correlation among the content as well.

Since our dataset was completely taken from the news domain, all the news documents had the published date as metadata. Moreover, in most cases, the same news document was published in all three languages on the same day. Therefore, before starting the aligning process, we filtered and grouped the documents using the published date and reduced the search space by a considerable amount.

We did the initial document alignment identification based on heuristics specific to the news website as described below.

[3] http://www.hirunews.lk.

[4] https://www.newsfirst.lk/.

[5] https://www.army.lk/.

[6] https://www.itnnews.lk.

- URL of each Hiru news document contains a unique id, which is shared by the news articles published in the respective languages. We used this property to identify candidate-aligned document pairs. These were varied by a human annotator, and the alignments accepted by the annotator were considered for the gold standard evaluation subset for Hiru News.
- Army news also had the publication date and time as shared attributes between the articles of the three languages. The same news was published in all three languages at the exact same date and time. Similar to Hiru news, we identified the candidate alignments for the Army news dataset using the publication date and time and later varied the alignment with the help of a human annotator.
- Documents crawled from NewsFirst and ITN websites did not have any such metadata that we could use to create the ground truth alignment. Therefore, ground truth alignment was manually created by human annotators, which was later varied by the same annotators by switching the datasets.

We consider the alignments verified by the annotators as the gold standard evaluation set. Altogether, eleven annotators were used to conduct the document alignment annotation.

The number of selected documents from each language pair along with the number of ground truth alignment pairs for each web source is shown in Table 1. Due to the low correlation between documents published by ITN, it has a lower number of aligned document pairs compared to other sources.

The aligned document pairs identified above were used as the input to the sentence alignment task. The number of input sentences on the source side and target side for each language pair is listed in Table 2. To conduct the sentence alignment annotation, we used five annotators altogether. Here also the annotations by one person were checked by another annotator for verification. Given a large number of sentences on each side, it would take a very long time for human annotators to find all sentence pairs that are translations of each other. Therefore, the gold standard sentence alignment evaluation set includes only 300 one-to-one sentence pairs from each website in all three language pairs.

Table 1 Statistics of document alignment evaluation dataset

Website	Sinhala–English			Tamil–English			Sinhala–Tamil		
	Si	En	Aligned	Ta	En	Aligned	Si	Ta	Aligned
Army	2033	2081	1848	1905	2081	1671	2033	1905	1578
Hiru	3133	1634	1397	2886	1634	1056	3133	2886	2002
ITN	6641	3212	1150	3035	3212	707	6641	3035	979
NewsFirst	3936	4273	1680	3929	3228	1266	3936	3929	1433

Table 2 Statistics of the sentence alignment evaluation dataset

Language pair	No of source sentences	No of target sentences
Sinhala–English	153,750	140,701
Tamil–English	87,266	87,330
Sinhala–Tamil	38,101	37,371

Table 3 Averages of the annotator scores for each label for document alignment and sentence alignment datasets

Task	En–Si		En–Ta		Si–Ta	
	CC	CB	CC	CB	CC	CB
Document alignment	77.67	22.33	75.33	24.67	84.00	16.00
Sentence alignment	97.33	2.67	78.00	22.00	77.67	22.33

3.2 Human evaluation on the benchmark evaluation datasets

On top of the annotation verification in the preceding stage, we have conducted a more systematic qualitative evaluation on the gold-standard dataset, following the methodology proposed by Kreutzer et al. [4]. From our gold standard evaluation set, we have sub-sampled 100 document pairs and sentence pairs from each language pair. Thereafter, we allocated three annotators to conduct the alignment verification per language pair, independently.

The annotation criteria are according to the work of Kreutzer et al. [4]. In their annotation scheme, the applicable labels for our dataset were CC (Correct translation, natural sentence) and CB (Correct translation, Boilerplate or low quality). When compiling our evaluation dataset, during the pre-processing stage we have already filtered out short sentences, so the label CS (Correct translation, Short) was not applicable. The rest of the annotation labels X (Incorrect translation, but both correct languages), WL (Source OR target wrong language, but both still linguistic content) and NL (Not a language: at least one of source and target are not linguistic content) were also not applicable in our case since the evaluation set had already undergone human annotation.

We calculated the number of annotations for each label given by each annotator and obtained the average scores as done by Kreutzer et al. [4]. The same approach was followed for annotating the sentence alignment samples as well. The outcome of the human evaluation for document alignment and sentence alignment samples are shown in Table 3.

More than 75% of document pairs have been annotated as correct alignments. When checked randomly, the ones that were considered as weak alignments had extra content on the target side which were not available in the source side and vice versa. In some other document pairs, the two languages had produced the same news incident from different perspectives. As a result, content-wise there were differences. This was another reason why some document pairs were annotated as CB.

For the sentence-aligned dataset, more than 77% had been annotated as correct alignments. However, for the En–Si language pair, the alignments were almost perfect, achieving an average alignment score of 97.33%. For the sentences marked as CB, we found that the target side had additional information compared to the source side and vice versa.

Therefore, based on the average scores we believe the gold standard evaluation set is fit to be recommended as a quality evaluation dataset.

3.3 Bilingual lexicons

As parallel data, we considered the bilingual lexicons: person names, designations, word dictionaries and glossaries. The English–Sinhala and English–Tamil Person Names and Designation lists were from the work of Priyadarshani et al. [63], while the Sinhala–Tamil bilingual lists were from Farhath et al. [64]. The bilingual dictionaries have been extracted

Table 4 Statistics of the bilingual lexicons

Bilingual lexicon	No of terms		
	Sinhala–English	Tamil–English	Sinhala–Tamil
Person names	6194	1374	76,334
Designations	6764	5779	44,193
Dictionary	23,722	36,551	19,132
Glossary	24,261	24,261	24,261

Table 5 Overview of the bilingual lexicons

	English-Sinhala		English-Tamil		Sinhala-Tamil	
Bilingual Lexicon	**En**	**Si**	**En**	**Ta**	**Si**	**Ta**
Person Names	ansha	අංශා	nalika	நாலிகா	තනුරාජ්	தனுராஜ்
	akila	අකිලා	ali	அலி	නිශාදි	நிஷாதி
Designations	operator	ක්‍රියාකරු	broker	தரகர்	සේවක	வேலையாள்
	major	මේජර්	mason	மேசன்	අංගණය	காலை
Dictionary	aback	පස්සට	the	என்ற	පාමුල	காலடி
	abed	ඇඳේහි	with	குல	පලතුර	பழம்
Glossary	abduction	අපහරණය	abduction	கடத்தல்	අපහරණය	கடத்தல்
	absent	අනුපස්ථිත	absent	வராத	අනුපස්ථිත	வராத

and used internally in an independent research and is yet to be published. A part of the Tamil–English dictionary is available at the WMT 2020 shared task [7]. We have obtained a Trilingual Glossary [8] from the Department of Official Languages, Sri Lanka. Statistics and samples of these bilingual lexicons are shown in Tables 4 and 5 respectively.

4 Methodology

In Sect. 4.1, we describe El-Kishky and Guzmán [13]'s method for document alignment, which is used as the baseline in this study, and our improvement as a weighting scheme considering bilingual lexicons. In Sect. 4.2, we describe the baseline system by Artetxe and Schwenk [25], followed by our improvement considering the bilingual lexicons.

4.1 Document alignment

Our document alignment system make use of multilingual sentence embeddings derived from PMLMs. In other words, we determine the alignment between two documents based on the semantic similarity between them, which is calculated by a distance scoring function. We use El-Kishky and Guzmán [13]'s technique as our baseline. We improve this distance scoring function, by introducing a weighting scheme using bilingual dictionaries.

[7] http://www.statmt.org/wmt20/translation-task.html.

[8] https://www.languagesdept.gov.lk/.

4.1.1 Baseline document alignment system

El-Kishky and Guzmán [13] defined a (1) distance scoring function to calculate the semantic distance between two documents and (2) a document matching algorithm to obtain the final aligned document pairs.

Distance scoring function Given a document pair, the objective of the distance scoring function is to calculate the semantic distance between two documents. If the semantic distance is less, then the degree of alignment increases. El-Kishky and Guzmán [13] introduced a novel distance metric named Cross-Lingual Sentence Mover's Distance (XLSMD). XLSMD was a distance metric based on Earth Mover's Distance (EMD). XLSMD represented each document as a normalized bag-of-sentences (nBOS) with all the sentences containing a pre-calculated probability mass (weight). Equation (1) shows the semantic distance between documents A and B. Here, $\Delta(i, j)$ is the Euclidean distance between the two sentences, which was calculated based on the sentence embeddings. As explained in Eq. (2), $T_{i,j}$ is how much of sentence i in document A was assigned to sentence j in document B (probability mass of a sentence).

$$XLSMD(A, B) = \min_{T \geq 0} \sum_{i=1}^{V} \sum_{j=1}^{V} T_{i,j} \times \Delta(i, j) \tag{1}$$

$$Subject\ to : \forall i \sum_{j=1}^{V} T_{i,j} = d_{A,i} \quad , \quad \forall j \sum_{i=1}^{V} T_{i,j} = d_{B,j} \tag{2}$$

Equation (3) shows the first function used for the probability mass calculation. Here, they used the relative frequencies of sentences as the probability mass. $\sum_{s \in A} count(s)$ represents the sentence count in document A. After calculating XLSMD, the distance was used in the document matching algorithm discussed next.

$$d_{A,i} = \frac{count(i)}{\sum_{s \in A} count(s)} \tag{3}$$

To make the XLSMD calculations more tractable, a greedy algorithm named Greedy Mover's Distance (GMD) an alternative to the relaxed-EMD was introduced. Here, the algorithm first calculated the Euclidean distance between each sentence pair and sorted them in ascending order. Then, it iteratively multiplies each distance by the smallest weight among the two sentences, which was named as the *flow* value as shown in Eq. (4).

$$distance = distance + \|s_A - s_B\| \times flow \tag{4}$$

However, Eq. (3) assigns probability mass uniformly across the sentences. Therefore, El-Kishky and Guzmán [13] introduced the following advanced weighting schemes in place of relative frequency.

Sentence length (SL) weighting

The SL weighting scheme was used under the assumption that longer sentences should be given more probability mass than shorter sentences. Equation (5) defines how this weight is calculated.

$$d_{A,i} = \frac{count(i) \times |i|}{\sum_{s \in A} count(s) \times |s|} \tag{5}$$

Here, $|i|$ and $|s|$ represent the number of tokens in the sentences i and s, respectively.

IDF weighting

IDF stands for inverse document frequency. Here, they have used the argument that the sentences that occur more frequently in the corpus should be given less importance than the infrequent sentences in the document. Equation (6) defines how it is calculated.

$$d_{A,i} = 1 + \log \frac{N+1}{1 + |d \in D : s \in d|} \tag{6}$$

Here, N is the total number of documents in domain D, and $|d \in D : s \in d|$ is the number of documents that contain sentence s.

SLIDF weighting

In this weighting scheme, both SL and IDF weights have been multiplied to obtain an aggregated weight as shown in Eq. (7). This was to give importance to both the number of tokens and the IDF of the sentence within the document collection.

$$d_{A,i} = SL(i) * IDF(i) \tag{7}$$

Similarly, the same weighting calculations SL, IDF and SLIDF had been done in the reverse direction, i.e., target to source document to calculate the probability mass for $d_{B,j}$ (i.e., j^{th} sentence in the target document B).

Document matching algorithm In this algorithm, initially, the semantic distances between each source document and target document were calculated according to the above-mentioned scoring function. Then, starting from the document pair containing the minimum distance, subsequent pairs d_A and d_B were selected iteratively, such that the documents d_A and d_B had not been considered in a previous selection.

4.1.2 New weighting scheme based on bilingual lexicons

As a novel contribution, we modify the distance scoring function of El-Kishky and Guzmán [13], by introducing a new weighting scheme considering bilingual lexicons. This weight calculation differs based on the nature of the term mapping in the bilingual lexicons (as word-to-word mappings or phrase-to-phrase mappings). This is described in the following section. With our improvement, the semantic distance calculation between a source side document d_A and target side document d_B is shown in Fig. 1.

We use the bilingual lexicons mentioned in Sect. 3.3 to introduce a weighting scheme on top of the SL, IDF and SLIDF schemes. Here, if a sentence s_A from document A contains a word in the bilingual lexicon and its translation in sentence s_B from document B, a variable *count* is incremented. The total of such words in sentence s_A is the final *count* value. The weighting between the two sentences s_A and s_B considering the variable *count* is shown in Eq. (8).

$$w_{A,B} = \frac{|s_A| - count}{|s_A|} \qquad |s_A| = Number\ of\ tokens\ in\ sentence\ s_A \tag{8}$$

The weighting $w_{A,B}$ is incorporated in to the GMD algorithm by modifying the distance calculation as shown in Eq. (9). Likewise, the distance is calculated considering each sentence pair in the two documents. Iterating through each sentence pair, the accumulated *distance* is the semantic distance between the document pair.

$$distance = distance + \|s_A - s_B\| \times flow \times w_{A,B} \tag{9}$$

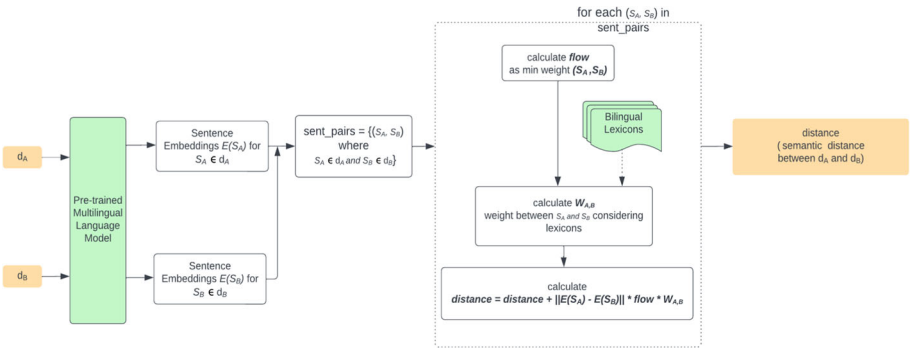


Fig. 1 Process for calculating the semantic distance between source language document d_A and target language document d_B. Here $w_{A,B}$ refers to the improved weight based on the bilingual lexicons. The accumulated *distance* scored from this process is the semantic distance between the document pair. Subsequently, the document matching algorithm (Sect.4.1.1) produces the final aligned document pairs

This way, when more words that map with the bilingual lexicons are identified in a sentence pair, the distance between the two sentences is lesser.

Usage of bilingual lists with one-word entries

Our person names list falls into this category. We added the parallel words in the person names bilingual list into a dictionary data structure where keys are words from language A and the values are arrays of translations of the key in language B. (Sometimes, one person's name has multiple translations due to multiple types of spelling formats.) When calculating the weights, for each sentence pair, we iterated through the words in the sentence to calculate the mapping counts. Here, we split the sentence s_A into words and check if each word w exists in the dictionary. If it exists, we get the parallel words v_B, and check if each parallel word exists in the sentence s_B. If so, we increase the counter and remove the mapped word from the sentence s_B. This counter value is used as the input in Eq. (8). Algorithm 1 in "Appendix A" explains this process.

Usage of bilingual lists with multi-word entries

Usage of designations bilingual list and word dictionary

Different to the person names bilingual lists, our designations and word dictionary fall into this category. Meaning, the entries contain more than one word (contains phrases). Therefore, when calculating weights, we implement a separate algorithm to identify the multiple word mapping considering the multiple words. Here, for each sentence s_A, we get all the permutations of words from length one to length five (the maximum length of a record in the dictionary is five). Then, we do the same process described above to get the mapping counts. Algorithm 2 depicts this process. When person names, designations, and word dictionaries are used in combination, we sum up the *count* values returned from both Algorithm 1 and 2 in "Appendix A", and use that value as the input for Eq. (8).

Improved dictionary

To improve the dictionary further, we add the terms from the glossary mentioned in Sect. 3.3. However, we could not see any improvement in terms of the scores. When investigated further, we observed that the glossary terms were mostly phrases and combined phrases as opposed to a single word. As a result, when the glossary was cross-checked with the sentence pairs, the number of overlapping terms was very low. Therefore, we utilized the parallel phrases in the dictionary to identify the distinct word pairs within the glossary terms. First, we cross-

Table 6 Overview of the improved dictionary

English-Sinhala		English-Tamil		Sinhala-Tamil	
En	Si	En	Ta	Si	Ta
horizontal	තිරස	zoned	வலயப்	ථොක්ක	குட்டு
horizontal	තිරස්	converging	குவிவு	සුම්	உரு அளவு
puffery	වංචනප්‍රාය ප්‍රචාරණය	workforce	வேலைப்படை	හිංසාව	காயம்

checked the phrases in the glossary with the words in the dictionary. We removed the parallel words that we found in the glossary phrases and extracted the remaining words from both languages as a parallel record. This way, the number of words in one record in the glossary got reduced by a considerable amount and we were able to improve the existing dictionary by adding the records we found from the glossary to the word dictionary. An overview of the improved dictionary is shown in Table 6.

4.2 Sentence alignment

Our sentence alignment system makes use of multilingual sentence embeddings derived from PMLMs. In other words, we determine the sentence alignment, considering the semantic similarity between the sentence pairs, calculated using these sentence embeddings. The baseline system is implemented according to the work of Artetxe and Schwenk [25]. Their method for semantic distance was defined as the margin-based cosine similarity over its nearest neighbours. We have introduced a weighting considering the bilingual lexicons, on top of this semantic distance.

4.2.1 Baseline sentence alignment system

Artetxe and Schwenk [25] obtained the LASER multilingual sentence embeddings for all the source and target sentences and aligned these sentence embeddings using a margin-based cosine similarity function. This similarity measurement considered a margin between the cosine of a given sentence pair and that of its respective nearest neighbours.

Artetxe and Schwenk [25] proposed the following three criteria for candidate generation, focusing a higher recall at the cost of precision.

- Forward: Each source sentence was aligned with exactly one best scoring target sentence. As a result, some target sentences may be aligned with multiple source sentences or with none.
- Backward: Equivalent to the forward strategy, but followed the candidate selection in the opposite direction.
- Intersection: Intersection of forward and backward criteria, with the objective of discarding inconsistent alignments.

4.2.2 Our improvements for sentence alignment

As our contribution to the sentence alignment task, we improve the semantic distance measurement step of Artetxe and Schwenk [25]'s method by introducing a weighting scheme using the bilingual lexicons (Sect. 3.3). When sentences in the source language document d_A

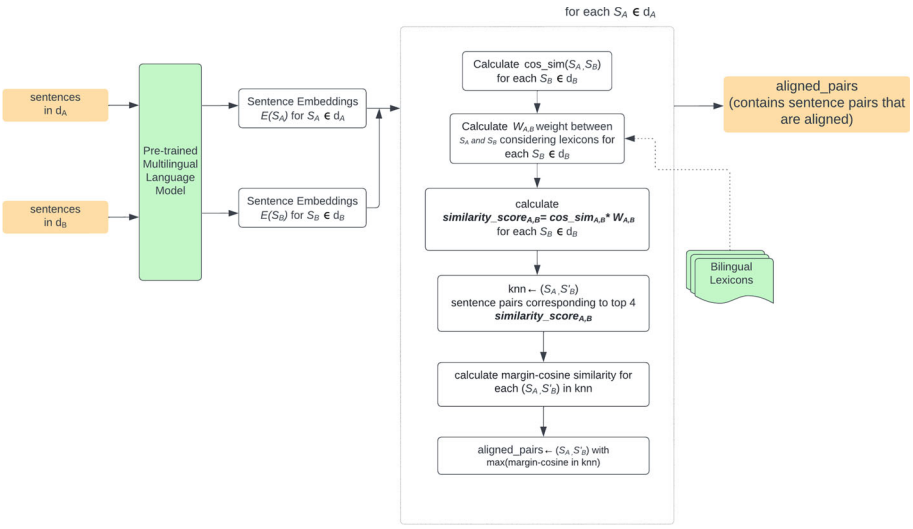


Fig. 2 This diagram outlines the sentence alignment algorithm for the forward criterion. Our baseline [25] uses margin-based cosine similarity as the semantic distance calculation function, while we improve this by introducing a weighting $w_{A,B}$, which is calculated using the bilingual lexicons. In the backward criterion, for each s_B in d_B, the aligned sentence is picked up from the source side. The semantic distance calculation is done in the reverse direction

and sentences in the target language document d_B are given as inputs, the sentence alignment algorithm produces the aligned parallel sentence pairs, as shown in Fig. 2.

In the forward criterion, based on the cosine similarity we select the best matching neighbourhood (k) of 4[9] candidates for each source sentence, similar to Artetxe and Schwenk [25]. Then, the margin-based cosine similarity is used over its nearest neighbours to determine the aligned target sentence. Here, if the source sentence s_A from document A contains a word w in the bilingual lexicon and the target sentence s_B from the selected k candidates contains the translation of the word w, the variable *count* is incremented. This *count* value is used to calculate the weight using Eq. (10) (multiplicative inverse of Eq. (8)), to give a higher weight for sentence pairs having more overlapping tokens and a lower weight for sentence pairs with a lower number of overlapping tokens.

$$w_{A,B} = \frac{|s_A|}{|s_A| - count} \qquad |s_A| = Number\ of\ tokens\ in\ source\ sentence\ s_A \tag{10}$$

New similarity score between each source sentence s_A and each target sentence s_B is calculated using Eq. (11), according to the selected k candidates.

$$similarity\_score_{A,B} = cosine\_similarity_{A,B} \times w_{A,B} \tag{11}$$

Then, each source sentence is aligned with the best scoring target sentence according to the above-calculated similarity scores.

In the backward criterion, for each sentence on the target side, an aligned sentence from the source side is identified. This is the reverse of the forward criterion method. Therefore, the weight calculation needs to be modified as shown in Eq. (12). Here, s_B refers to the selected sentence from the target side, $w_{A,B}$ refers to the weight between s_B and the nearest

[9] We use k = 4 for all experiments in this work as it gave the best results in all our experiments.

neighbours identified from the source side. The *count* is incremented when a word in s_B exists in the bilingual lexicon as well as in the source sentence retrieved from nearest neighbours. The nearest neighbour retrieval is based on the cosine similarity, similar to Artetxe and Schwenk [25].

$$w_{B,A} = \frac{|s_B|}{|s_B| - count} \qquad |s_B| = Number\ of\ tokens\ in\ source\ sentence\ s_B \tag{12}$$

The final similarity score between sentence s_B and s_A is shown in Eq. (13)

$$similarity\_score_{B,A} = cosine\_similarity_{B,A} \times w_{B,A} \tag{13}$$

In the intersection criterion, the intersection of the sentence pairs identified from the forward criterion and backward criterion are taken. Therefore, this is identical to the work by Artetxe and Schwenk [25]

5 Evaluation

We evaluated our improvements separately for document alignment and sentence alignment tasks, using the golden alignment dataset we prepared (see Sect. 3). Further, an extrinsic evaluation was conducted on sentence alignment by training an NMT system.

5.1 Document alignment

El-Kishky and Guzmán [13] used LASER multilingual sentence embeddings in their experiments. Therefore, for document alignment task, we report the results for the baseline system only using LASER embeddings. For search efficiency, Subsequently, an ablation study is conducted by sequentially adding each bilingual lexicon on top of the previous experiment. Then, we repeat the above experiments for XLM-R and LaBSE. Thus, this becomes the first empirical study of these three models for the task of document alignment.

Similar to El-Kishky and Guzmán [13], our technique is aimed at high recall at the cost of low precision. However, we have reported the recall (R), precision (P) and $F1$ scores over the gold-standard evaluation set. We experimented with English–Sinhala, English–Tamil and Sinhala–Tamil language pairs for each news web source. For each language pair, we report the averages of the individual scores obtained for the news sources in Table 7. Results per news source are reported in Appendix B1

The document alignment results show that the baseline result of El-Kishky and Guzmán [13] has been outperformed by our improvement when incorporating all the bilingual lexicons (BL+N+Ds+MDc) for all three language pairs. Considering the averaged $F1$ scores, the improvement is significant (around 44% increase compared to the baseline) for the En–Ta language pair. The improvement for Si–Ta is 13% and for En–Si 2%, respectively. This is an interesting observation. Sinhala and Tamil are considered to be under-represented in LASER, meaning that the cross-lingual alignment related to these languages is weak. Therefore, by using bilingual dictionaries can enhance the cross-lingual alignment between language pairs. Additionally, the performance of document alignment depends on the correlation between source and target documents. One good example is the Army news source, on which even the baseline system performed well.

Further, it was noted that the results for En–Ta were very low compared to the other language pairs, En–Si and Si–Ta. Tamil belongs to the Dravidian family, and this language

Table 7 Averaged recall (R), precision (P) and $F1$ scores obtained for each news source with respective to the language pairs

Experiment	Weighting	Averaged Scores across each News Source								
		En–Si			En–Ta			Si–Ta		
		R	P	$F1$	R	P	$F1$	R	P	$F1$
LASER										
BL	SL	92.22	62.56	72.38	50.47	31.50	37.28	73.41	40.80	50.23
	IDF	89.97	60.95	70.55	48.06	29.29	34.96	71.04	39.05	48.26
	SLIDF	92.21	62.56	72.38	50.64	31.59	37.40	73.24	40.75	50.15
BL+N	SL	93.30	63.28	73.22	51.58	32.22	38.12	77.41	43.28	53.18
	IDF	91.14	61.81	71.52	49.08	29.89	35.70	75.27	41.52	51.26
	SLIDF	93.32	63.29	73.23	51.71	32.30	38.22	77.45	43.33	53.23
BL+N+Ds	SL	93.30	63.28	73.22	51.58	32.22	38.12	77.41	43.28	53.18
	IDF	91.14	61.81	71.52	49.08	29.89	35.70	72.38	39.51	48.94
	SLIDF	93.32	63.29	73.23	51.71	32.30	38.22	77.45	43.33	53.23
BL+N+Ds+Dc	SL	93.66	63.54	73.51	70.74	42.87	51.39	79.61	44.46	54.67
	IDF	91.14	61.81	71.52	67.19	40.33	48.52	77.53	42.87	52.89
	SLIDF	93.32	63.29	73.23	70.64	42.82	51.33	79.61	44.46	54.67
BL+N+Ds+MDc	SL	**94.14**	**63.78**	**73.83**	73.84	44.70	53.62	**82.99**	**46.46**	**57.09**
	IDF	91.14	61.81	71.52	69.93	41.90	50.45	80.51	44.64	55.04
	SLIDF	93.32	63.29	73.23	**73.94**	**44.74**	**53.68**	82.95	46.43	57.06

Table 7 continued

Experiment	Weighting	Averaged Scores across each News Source								
		En–Si			En–Ta			Si–Ta		
		R	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1
XLM-R										
BL	SL	96.75	65.47	75.82	92.12	55.45	66.80	92.18	52.17	63.95
	IDF	96.66	65.42	75.76	92.31	55.54	66.91	91.05	51.21	62.90
	SLIDF	96.71	65.44	75.79	92.11	55.45	66.80	92.19	52.18	63.96
BL+N	SL	97.33	65.95	76.34	92.51	55.75	67.13	93.03	52.80	64.66
	IDF	96.84	65.63	75.96	91.86	55.42	66.71	92.00	51.98	63.76
	SLIDF	97.37	65.98	76.38	92.50	55.74	67.13	93.02	52.80	64.65
BL+N+Ds	SL	97.33	65.95	76.34	92.51	55.75	67.13	93.03	52.80	64.66
	IDF	96.84	65.63	75.96	91.86	55.42	66.71	91.17	51.25	62.97
	SLIDF	97.37	65.98	76.38	92.50	55.74	67.13	93.02	52.80	64.65
BL+N+Ds+Dc	SL	97.28	65.91	76.30	94.37	56.90	68.51	93.31	52.96	64.86
	IDF	96.84	65.63	75.96	93.37	56.28	67.77	92.13	52.08	63.87
	SLIDF	97.37	65.98	76.38	94.39	56.91	68.52	93.33	52.97	64.87
BL+N+Ds+MDc	SL	**97.56**	**66.14**	**76.55**	**94.48**	**56.97**	**68.59**	**93.77**	**53.30**	**65.24**
	IDF	96.84	65.63	75.96	93.42	56.27	67.77	92.57	52.47	64.28
	SLIDF	97.37	65.98	76.38	94.41	56.32	67.91	93.74	53.28	65.22

Table 7 continued

Experiment	Weighting	Averaged Scores across each News Source								
		En–Si			En–Ta			Si–Ta		
		R	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1
LaBSE										
BL	SL	98.26	66.64	77.12	95.81	57.76	69.56	96.35	55.37	67.55
	IDF	98.18	66.61	77.08	95.40	57.46	69.21	96.46	55.37	67.56
	SLIDF	98.24	66.62	77.11	**95.42**	**58.24**	**69.99**	96.46	55.38	67.57
BL+N	SL	98.25	66.63	77.12	95.67	57.67	69.45	96.30	55.33	67.50
	IDF	98.20	66.64	77.11	94.95	57.28	68.96	96.11	55.16	67.31
	SLIDF	98.25	66.63	77.12	95.65	57.66	69.43	96.34	55.36	67.54
BL+N+Ds	SL	98.25	66.63	77.12	95.67	57.67	69.45	96.30	55.33	67.50
	IDF	98.20	66.64	77.11	94.95	57.28	68.96	96.48	55.38	67.58
	SLIDF	98.25	66.63	77.12	95.65	57.66	69.43	96.34	55.36	67.54
BL+N+Ds+Dc	SL	98.27	66.63	77.11	94.99	57.27	68.96	96.51	55.41	67.61
	IDF	98.20	66.64	77.11	95.66	57.64	69.42	96.36	55.27	67.46
	SLIDF	98.25	66.63	77.12	95.64	58.42	69.82	96.53	55.42	67.62
BL+N+Ds+MDc	SL	**98.31**	**66.67**	**77.16**	95.64	58.42	69.82	**96.58**	**55.45**	**67.66**
	IDF	98.20	66.64	77.11	95.08	57.32	69.02	96.17	55.15	67.31
	SLIDF	98.25	66.63	77.12	95.57	57.58	69.35	96.51	55.43	67.62

Here, BL refers to the recreated baseline [13] considering LASER embeddings. On top of this, each bilingual lexicon had been added and the experiments were repeated. Bilingual lexicons: (Sect. 3.3) Person Names (N), Designations (Ds), Dictionary (Dc) and Improved Dictionary (MDc). Subsequently, considering XLM-R and LaBSE, the same set of experiments was repeated

family is under-represented in many pre-trained models. Moreover, Dravidian languages have a higher linguistic distance from English, compared to the Indo-Aryan family, to which Sinhala belongs. We suspect these are the reasons to produce a lower result for the En–Ta language pair.

Both XLM-R and LaBSE outperformed the LASER scores. A further observation was that the LaBSE baseline was higher than that of XLM-R. XLM-R and LaBSE had been pre-trained using a massive collection of monolingual data using the transformer architecture, while LASER was built on the RNN architecture. Hence, we believe that they have captured the cross-lingual features better than LASER. Additionally, LaBSE had also used parallel data to improve the multilingual embeddings and to strengthen the cross-lingual transfer. As a result, LaBSE embeddings are more favourable for the document alignment task.

XLM-R and LaBSE baseline scores being better than the LASER scores for all three language pairs suggest that the multilingual embeddings obtained via self-supervised learning have a better language representation for low-resource languages, compared to LASER, which was trained in a supervised manner. This is very beneficial for non-English centric language pairs such as Si–Ta, which have been explored to a lesser extent.

When it comes to XLM-R results, the absolute average $F1$ score gains produced by using lexicons are in the range of 0.7 for En–Si, 1.7 for En–Ta and 1.3 points for Si–Ta, respectively. When LaBSE is used, these gains are less than 0.5 $F1$ points. Therefore, we can conclude that XLM-R and LaBSE already have rich cross-lingual alignment information, and the amount of additional information provided by bilingual lexicons is relatively less.

Considering the experiment using LASER embeddings, we could observe that the gains were maximum when using all bilingual lexicons. Therefore, if we could find more lexicons we could increase the task performance. Even though the person names bilingual list of Sinhala–Tamil is about ten times larger than that for the other language pairs, we could not see a considerable improvement in Sinhala–Tamil compared to the other two. This may be due to the inflected nature of the two languages. The names could be in the inflected form in the parallel content, while the lexicons contain the names in the base form.

5.2 Sentence alignment

For the sentence alignment experiments, we used three baselines:

1. Artetxe and Schwenk [25]'s method. As mentioned in Sect. 4.2, they used LASER multilingual embeddings and considered the alignments based on Forward, Backward and Intersection criteria using margin-based cosine similarity as the distance calculation method.
2. Hunalign [14], for the purpose of comparing our work with a statistical method. Hunalign has been used as a baseline for other research that experimented with embedding-based techniques for sentence alignment [5]
3. Feng et al. [18]'s method. They conducted sentence alignment using raw cosine similarity over the sentence embeddings obtained from LaBSE. This baseline is useful to compare the effect of the margin-based cosine similarity [25] and raw cosine [18] distance measurement.

We applied our improvement to Artetxe and Schwenk [25]'s method, using LASER embeddings, as done by Artetxe and Schwenk [25]. Then, these experiments were repeated for XLM-R and LaBSE. We conducted the dictionary improvement on top of Feng et al. [18]'s baseline as well.

As our ground-truth alignment contain only a small fraction (approx. 300) of parallel sentences, there can be many more valid cross-lingual sentence pairs in these datasets. Therefore, we evaluated aligned sentence pairs using recall (i.e., what percentage of the sentence pairs in the golden alignment set are found by the algorithm), which was one of the commonly used measurements in other research as well [25, 65]. The results are shown in Table 8. Note that since the use of all the bilingual lexicons gave the best result for the document alignment task, here we have considered all the bilingual lexicons for the BL+Dic (baseline with dictionary improvement) experiments.

First and foremost, we note that multilingual embedding-based methods significantly outperform Hunalign [14]. Even the baseline [25] outperforms Hunalign by a significant margin of 74% with respect to recall for En–Si languages.

Compared to the LaBSE baseline that uses raw cosine similarity [18], Artetxe and Schwenk [25]'s margin-based cosine similarity reports a recall value that is around 3% higher for the Sinhala–Tamil and Tamil–English pairs. Therefore, we can conclude that margin-based cosine similarly is favourable for the sentence alignment task.

Our sentence alignment system that incorporates bilingual lexicons outperforms Artetxe and Schwenk [25]'s method in all three language pairs for all the websites with the exception of very few as seen in Table 8. Tamil–English language pair shows the highest improvement by outperforming the baseline system by on average 15%. For Sinhala–Tamil and Sinhala–English pairs, on average 8% and 4% recall gains (respectively) were obtained for LASER embeddings.

Baseline sentence alignment results for both Tamil–English and Sinhala–Tamil language pairs are considerably low compared to Sinhala–English for LASER embeddings. The low amount of training data used for Sinhala and Tamil when training the LASER toolkit could be the reason for that [19]. Further, Tamil belongs to the Dravidian family, and this language family is under-represented in many pre-trained models. Moreover, Dravidian languages have a higher linguistic distance from English, compared to the Indo-Aryan family, to which Sinhala belongs. We suspect these are the reasons to produce a lower result for the En–Ta language pair. This is the same observation with the document alignment results as well.

The bilingual lexicon terms are in nominative form. However, Sinhala and Tamil are morphologically rich languages, which means words are inflected based on gender, plurality, or morphological case category. Although the word may exist in nominative form in the bilingual dictionary, in the sentences they can be in the inflected form. So the dictionary improvement fails to identify such cases. We suspect this as a main reason for our improvement to be marginal specific to Si–Ta language pair.

We observe that the baseline sentence alignment scores considering XLM-R and LaBSE have outperformed the LASER baseline. Further, the LaBSE baselines produce the highest scores across all three language pairs. XLM-R had been trained on massive collection of monolingual data, while LaBSE had been pre-trained using monolingual data and fine-tuned using parallel data. The underlying reason for the improvement in scores we believe is the improvement in the language representations. Although XLM-R had been purely on unsupervised manner, it had still managed to capture cross-lingual features in the languages to be favorable for the sentence alignment task. Since LaBSE had been fine-tuned using parallel data, we experience that this step had helped to improve the cross-lingual alignments further, in the embeddings produced by the model. As a result, the LaBSE scores are the highest.

According to the results, our dictionary improvement is not that much significant compared to XLM-R and LaBSE baselines for Si–En and Ta–En language pairs. For Si–Ta, our improvement produces a gain of + 0.5 recall. This shows that the multilingual representations

Table 8 Sentence alignment results in terms of recall (R)

PMLM	Exp.	Army			Hiru			ITN			Newsfirst			Averaged Scores					
														Forward		Backward		Intersection	
		FW	BW	INT	FW	BW	INT	FW	BW	INT	FW	BW	INT	Para Sents	R	Para Sents	R	Para Sents	R
Sinhala–English																			
Hugalign [14]		29.00			11.63			9.00			11.00			7266	15.16				
LaBSE [18]	BL	98.67	97.34	98.33	**97.34**	**97.34**	**97.34**	99.33	**99.33**	99.00	98.33	98.67	**98.33**	23847	98.42	23385	98.17	18705	98.25
	BL+Dict	98.00	98.00	97.00	**97.34**	**97.34**	**97.34**	98.33	99.00	99.00	98.33	**99.00**	**98.33**	23847	98.00	23385	98.34	17848	97.92
LASER [25]	BL	94.33	97.00	93.33	95.35	95.35	94.02	93.33	***93.33***	89.33	95.67	94.33	92.33	23847	94.67	23385	95.00	17439	92.25
	BL+Dict	96.33	97.33	94.33	95.68	95.68	94.35	95.33	97.00	94.00	97.67	96.33	96.00	23847	96.25	23385	96.59	17694	94.67
XLM-R	BL	92.33	93.33	89.67	96.35	96.68	95.68	94.00	96.00	92.33	96.67	95.33	94.33	23847	94.84	23385	95.34	17726	93.00
	BL+Dict	96.00	94.67	93.00	**97.34**	96.68	96.68	96.67	96.67	96.00	97.33	96.67	95.67	23847	96.84	23385	96.17	18079	95.34
LaBSE	BL	**99.00**	**99.33**	**99.00**	**97.34**	**97.34**	**97.34**	**99.67**	**99.33**	**99.33**	98.33	98.67	**98.33**	23847	98.59	23385	98.67	18684	**98.50**
	BL+Dict	**99.00**	**99.33**	**99.00**	**97.34**	**97.34**	**97.34**	**99.67**	**99.33**	**99.33**	**98.67**	**99.00**	**98.33**	23847	**98.67**	23385	**98.75**	18665	**98.50**
Tamil–English																			
LaBSE [18]	BL	94.67	93.33	89.33	88.67	85.33	80.33	90.60	90.60	89.60	96.00	95.67	95.67	18242	92.48	18606	91.23	12762	88.73
	BL+Dict	93.33	94.67	92.33	88.00	**86.33**	81.67	91.61	**91.95**	88.59	95.67	**96.33**	94.67	18242	92.15	18606	92.32	13647	89.31
LASER [25]	BL	77.33	73.67	67.67	68.00	52.00	44.33	67.11	62.75	54.03	74.33	65.33	60.33	18242	71.69	18606	63.44	9925	56.59
	BL+Dict	84.67	80.67	76.00	78.67	61.67	56.33	80.54	73.83	69.13	85.33	76.00	73.67	18242	82.30	18606	73.04	10984	68.78
XLM-R	BL	86.67	88.33	82.00	83.00	78.33	72.67	83.22	83.56	78.86	92.33	91.33	89.33	18242	86.31	18606	85.39	12643	80.71
	BL+Dict	88.33	91.33	84.00	83.67	79.67	74.67	85.91	84.56	82.22	92.67	93.00	91.00	18242	87.64	18606	87.14	12974	82.97
LaBSE	BL	**96.33**	96.33	94.67	**89.67**	**86.33**	**83.33**	**92.62**	**91.95**	**91.28**	96.33	**96.33**	96.00	18242	**93.74**	18606	92.74	13699	**91.32**
	BL+Dict	**96.33**	**97.00**	**95.33**	88.33	**86.33**	82.33	92.28	91.61	90.60	**96.67**	**96.33**	**96.33**	18242	93.40	18606	**92.82**	13647	91.15

Table 8 continued

PMLM	Exp.	Army			Hiru			ITN			Newsfirst			Averaged Scores					
														Forward		Backward		Intersection	
		FW	BW	INT	FW	BW	INT	FW	BW	INT	FW	BW	INT	Para Sents	*R*	Para Sents	*R*	Para Sents	*R*
Sinhala–Tamil																			
LaBSE [18]	BL	93.38	93.38	90.73	93.38	93.38	90.73	97.00	97.33	95.67	96.00	97.33	93.33	21041	94.94	20642	95.35	13145	92.61
	BL+Dict	93.71	91.72	89.73	96.33	97.00	94.67	99.00	95.67	95.00	97.00	96.33	**96.67**	21041	96.51	20642	95.18	13693	94.02
LASER [25]	BL	71.52	79.47	66.56	75.00	80.33	69.00	73.00	81.33	65.33	73.00	83.00	67.33	21041	73.13	20642	81.03	11652	67.06
	BL+Dict	74.50	81.46	69.54	80.33	88.00	76.00	81.33	86.00	71.67	78.00	88.67	70.33	21041	78.54	20642	86.03	12051	71.88
XLM-R	BL	83.44	81.46	78.15	90.67	91.00	87.33	91.33	90.00	87.00	93.67	95.33	92.33	21041	89.78	20642	89.45	13989	86.20
	BL+Dict	86.09	82.45	79.47	92.00	94.33	91.00	93.33	93.33	90.00	95.00	**98.67**	93.00	21041	91.61	20642	92.20	14120	88.37
LaBSE	BL	**95.03**	**94.70**	**92.38**	97.33	**98.00**	96.33	99.33	**98.67**	**98.33**	**98.67**	97.67	95.33	21041	**97.59**	20642	97.26	14853	95.60
	BL+Dict	**95.03**	**94.70**	**92.38**	**97.67**	**98.00**	**97.00**	**99.67**	**98.67**	**98.33**	97.00	**98.67**	**96.67**	21041	97.34	20642	**97.51**	14889	**96.10**

Hungalign [14] and LaBSE [18] refer to the re-created baseline scores for our dataset. Here, BL refers to the score obtained using Artetxe and Schwenk [25]'s method, and BL+Dic refers to the scores obtained using our bilingual lexicon improvement. The experiments have been conducted considering the Forward (FW), Backward (BW) and Intersection (INT) criterion

of XLM-R and LaBSE have frame for improvement when it comes to non-English centric diverse language families such as Sinhala and Tamil.

5.3 Extrinsic evaluation with NMT

To analyse the effectiveness of incorporating bilingual lists and different multilingual embeddings into the sentence alignment task, we conducted an extrinsic evaluation by training NMT systems with the obtained parallel sentences. We merged the parallel sentences obtained from each news source and trained NMT systems specific for the language pair in the forward and in reverse directions.

We used the SiTa trilingual (Sinhala, Tamil and English) parallel machine translation (MT) evaluation sets [66] created by the National Languages Processing Center of the University of Moratuwa, Sri Lanka[10] to evaluate the NMT performance. Additionally, we report the MT scores for the Flores v1 [67] evaluation set for Sinhala–English and Flores-101 [68] multilingual evaluation set for Tamil–English language pairs.

More recently, NMT systems fine-tuned on the mBART50 sequence-to-sequence pre-trained model [69] had been successful in terms of Sinhala and Tamil [70, 71]. Therefore, in order to build an NMT model, we decided to fine-tune the mBART50 model with the parallel sentences obtained from the sentence alignment task. Experiments were done using the fairseq toolkit [72], and the performance was evaluated using the evaluation datasets mentioned above. BLEU scores were obtained with sacreBLEU [73].

The NMT results shown in Table 9 are rather low, which we believe is due to the following reasons: (1): the SiTa evaluation dataset has been obtained from the official document domain, while the Flores evaluation datasets have been obtained from Wikipedia. In contrast, we mined the parallel corpus from the news domain. Therefore, the domain difference is identified as the primary reason for the NMT systems to produce low results. (2) The parallel corpus size produced by the sentence alignment task is in the range of 9,000-23000, which marks an extremely low-resource setting [3]. Both these reasons lead to the NMT system producing a low result. However, we believe that this is not a bottleneck in conducting our study as we are only interested in analysing the impact of the bilingual lexicon integration on the sentence alignment task.

We observe that comparable results are obtained across all languages for Backward and Intersection criteria for NMT models for Si→En, Ta→En and Si→Ta. In the backward criterion, for each target language sentence, an aligned sentence from the source language is obtained. Therefore, the selected source sentence might not always guarantee a proper translation for the target sentence. This can be identified as a weak parallel sentence pair with the noise at the source side. This is an interesting observation as it indicates that the NMT is robust to source side noise. However, when the noise is in the target side (as in the case of Forward criterion), it degrades the performance of the NMT. Since the Intersection is dependent on the Backward criterion, the improvement can also be seen in NMT systems trained with the Intersection criterion. In the NMT systems trained for En→Si, En→Ta and Ta→Si, the same observation is true for Forward and Intersection criteria. Here, the target language for the NMT system is picked up from the forward criterion. That is, in the case of En→Si NMT, with the Forward criterion, for each Si sentence, an En sentence is identified. So here the noisy sentence is found on the source-side (En). Therefore, for the NMT systems in the reverse direction, the Forward criterion is favourable.

[10] https://uom.lk/nlp

Table 9 BLEU scores for NMT systems trained with parallel data obtained from sentence alignment step considering Forward (F), Backward (B) and Intersection (I) criterion

PMLM	Exp.	F		B		I		F		B		I		F	B	I
		Si→En						Ta→En						Si→Ta		
		ST	FL	ST	FL	ST	FL	ST	FL	ST	FL	ST	FL	ST	ST	ST
LASER	BL	9.7	3.9	11.6	5.6	12.0	6.3	3.8	2.1	6.4	4.1	6.6	4.8	3.5	4.4	4.5
	BL+Dict	**9.9**	**4.4**	12.2	**6.6**	**12.4**	6.4	**5.5**	4.3	7.7	5.5	7.3	5.1	3.8	4.9	4.6
XLM-R	BL	8.8	4.0	11.4	5.6	11.9	6.5	4.0	4.1	6.1	5.1	7.7	5.9	3.7	4.1	4.7
	BL+Dict	9.0	3.6	11.8	6.0	12.1	6.4	4.6	**5.5**	7.0	5.4	7.7	5.8	3.9	4.7	4.6
LaBSE	BL	9.5	4.3	11.9	6.3	11.9	**6.6**	3.8	4.4	8.1	5.8	8.2	6.2	**4.0**	4.7	4.7
	BL+Dict	9.3	4.1	**12.4**	6.5	12.1	**6.6**	3.9	5.4	**8.2**	**6.3**	**8.4**	**6.5**	**4.0**	**5.2**	**4.9**
		En→Si						En→Ta						Ta→Si		
LASER	BL	**8.3**	**1.8**	6.5	0.6	8.5	1.4	4.5	1.3	3.8	0.5	4.4	0.7	4.8	3.1	6.4
	BL+Dict	**8.3**	1.6	6.9	0.5	8.6	1.5	4.5	**1.5**	4.1	0.7	4.4	1.1	6.6	3.3	**6.5**
XLM-R	BL	8.0	1.7	7.0	0.6	7.9	1.7	4.6	1.3	4.2	0.9	4.4	**1.4**	5.6	**4.6**	6.1
	BL+Dict	8.1	**1.8**	**7.9**	**0.8**	8.3	1.8	**4.7**	1.4	4.1	0.9	4.5	1.3	5.9	4.3	5.7
LaBSE	BL	8.2	1.7	7.4	**0.8**	8.2	**2.0**	**4.7**	1.1	**4.3**	0.8	4.6	1.2	**6.9**	4.4	6.1
	BL+Dict	8.2	1.7	7.2	**0.8**	**8.7**	1.9	4.5	1.4	4.2	**1.0**	**5.0**	**1.4**	5.9	4.3	6.4

The results have been reported against the SiTa (ST) and Flores (FL) evaluation testsets

We see that the NMT scores obtained by bilingual lists have improved over the baseline scores for most of the cases as per Table 9. This means that bilingual list integration has improved the quality of the parallel sentences. Considering the SiTa evaluation set, the maximum gain provided for LASER is +1.8 BLEU, XLM-R is +0.9 BLEU and for LaBSE it is +0.5 BLEU. Similarly, for Flores evaluation set, it is +1.4, +0.6 and +0.5 BLEU for LASER, XLM-R and LaBSE (respectively). Here we can see identical patterns with respect to both evaluation sets. The gain is the highest for LASER while for XLM-R and LaBSE it is in the same range. Although the Wikipedia data have been used during training these multilingual PMLMs, it is evident that the multilingual embeddings are not biased to the evaluation set on Wikipedia.

For Ta→En and Ta→Si directions, it shows a maximum improvement of +1.7 BLEU and +1.3 BLEU scores (respectively) for the LASER embeddings for the SiTa evaluation set. As Tamil is an under-represented language in the LASER training data, the lexicon integration has managed to improve the NMT scores.

In sentence alignment results, the scores were always in increasing order for LASER, XLM-R and LaBSE, respectively. However, for the downstream NMT task, we observed that the scores were mostly high for LASER and LaBSE compared to XLM-R. Although we expected the sentence alignment scores and NMT scores to follow the same pattern, it was not the case. LASER had been trained purely on parallel data while LaBSE had been pre-trained using monolingual and parallel data, followed by a fine-tuning phase with parallel data. Therefore, we observe that multilingual systems pre-trained with parallel data perform better in the NMT downstream task.

6 Further analysis

We conducted further analysis to identify the impact of lexicon integration on sentence alignment. Table 10 shows three scenarios where lexicon integration did not work. An example is given for the Sinhala–English pair. However, these findings are valid for other language pairs as well.

As explained by scenario A, the sentence pair that should be aligned does not contain any overlapping terms with the bilingual lexicons. Hence, it cannot be benefited by our lexicon integration. Further, the En sentence and another Si sentence from the same context have overlaps in terms of parallel lexicons. As a result, the sentence alignment algorithm selects an incorrect Sinhala sentence as the alignment for the English source sentence.

In scenario B, when there are equal overlaps between the candidate aligned sentences, the lexicon improvement is not effective. In such instances, the alignment is purely determined by the margin-based cosine similarity. In this example, both Sinhala candidate sentences have two lexicon overlaps; therefore, the selection of the aligned sentence cannot be based on the integrated lexicon.

According to scenario C, the sentences contain lexicon terms, but in an inflected form. Thus, our algorithms cannot identify those lexical terms appearing in sentences. In the example, the lexicon overlaps are missed for two word-pairs owing to inflections (in both En and Si). If the inflections were accounted in the algorithm, the correct alignment sentence-pair could be identified. We believe if a matching can be done at the lemma, a further improvement can be obtained. However, for Sinhala and Tamil, there is no lemmatizer that guarantees the coverage of the full vocabulary. Therefore, at present, working at the lemma level is not feasible.

Table 10 Error analysis in the sentence alignment task. Here, the alignment[corr] refers to the alignment in the gold-standard evaluation set and alignment[incorr] refers to the alignment produced in the experiments

Scenario	Correct/ Incorrect	Example
A	alignment [corr]	ජාතික දිනයේ දී ජාතික ගීය ගායනා කර රට වෙනුවෙන් දිවිපිදූ රණවිරුවන් සිහිපත් කරමින් විනාඩි දෙකක නිශ්ශබ්දතාවයක් ආරක්ෂා කිරීමෙන් අනතුරුව මුලිකව් ආරක්ෂක සේනා ආඥාපති මේජර් ජෙනරාල් දුෂ්යන්ත රාජගුරු ගේ ප්‍රධානත්වයෙන් සැමරුම් උත්සව කටයුතු ආරම්භ කරන ලදී.
		Major General Dushyantha Rajaguru, Commander, Security Forces - Mullaittivu early morning on the National Day, began commemorative proceedings with the singing of the National Anthem and observance of a two-minute silence in memory of all fallen War Heroes.
	Lexicon overlap	No overlaps
	alignment [incorr]	එමෙන්ම, කිලිනොච්චි ආරක්ෂක සේනා ආඥාපති මේජර් ජෙනරාල් යුල්ෆ් නුගේරා ගේ උපදෙස් මත 71 වන ජාතික නිදහස් දිනයට සමගාමීව පෙබරවාරි මස 3 සහ 4 දිනයන්හි තවත් ප්‍රජා සන්නාරක ව්‍යාපෘතීන් කිහිපයක් කිලිනොච්චි ආරක්ෂක සේනා මූලස්ථානය විසින් දියත් කරන ලදී.
		Major General Dushyantha Rajaguru, Commander, Security Forces - Mullaittivu early morning on the National Day, began commemorative proceedings with the singing of the National Anthem and observance of a two-minute silence in memory of all fallen War Heroes.
	Lexicon overlap	{'සහ': 'and', 'මත': 'on'}
B	alignment [incorr]	එහිදී 122 වන බලසේනාවේ බලසේනාධිපති කර්නල් මොහාන් රත්නායක විසින් බලප්‍රදේශයේ යුද්ධ හමුදා සාමාජිකයින් සිදුකරනු ලබන කාර්යභාර්ය පිළිබඳ ආඥාපතිතුමන් දැනුවත් කළහ.
		Afterwards, he was accorded a Guard of Honour by troops of 18 Gemunu Watch in conformity with military traditions.
	Lexicon overlap	{'පිළිබඳ': 'in', 'හමුදා': 'military'}
	alignment [corr]	ඉන් අනතුරුව, 18 වන ගැමුණු මේවා බලකායේ භට පිරිස් විසින් පිරිනමනු ලැබූ හමුදා සම්ප්‍රදායානුකූල සම්මාන ආචාර පෙළපාලියේ ගෞරවාචාරය ද එතුමන් වෙත පිරිනැමීය.
		Afterwards, he was accorded a Guard of Honour by troops of 18 Gemunu Watch in conformity with military traditions.
	Lexicon overlap	{'මේවා': 'military', 'හමුදා': 'military'}
C	alignment [incorr]	මෙහිදී ධජය එසවීම, ජාතික ගීය සහ යුද්ධ හමුදා ගීතය ගායනා කිරීම, රාජ්‍ය සේවයේ කැපවීම පිළිබඳව ප්‍රතිඥාව කියවීම, මියගිය රණවිරුවන් සිහිපත් කිරීම සඳහා විනාඩි 2 ක නිහඬතාවයක් පැවැත්වීම මෙන්ම නව බලාපොරොත්තුව හා සාර්ථකත්වය පිළිබඳ විස්තර කරමින් නව වසරේ වැඩ ඇරඹීමට සුභ පැතුම් එක්කරමින් යුද්ධ හමුදාධිපතිතුමන් විසින් නව වසර සඳහා නිකුත් කරන ලද පණිවිඩය කියවනු ලැබිය.
		Similarly, strict disciplinary action should be taken against any violators of discipline and this should be borne in your mind all the time," the Commander warned during his speech to the troops at the SLCMP Headquarters.
	Lexicon overlap	{'හා': 'and', 'සහ': 'and'}
	alignment [corr]	ජාතික ධජය <u>එසවීම</u>,රාජ්‍ය ප්‍රතිඥාව සහ මියගිය රණවිරුවන් සිහිකිරීම සඳහා විනාඩි දෙකක <u>නිශ්ශබ්දතාවයක්</u> පැවැත්වීමෙන් පසු නව වසරේ රාජකාරී ආරම්භ කරන ලදී.
		<u>Hoisting</u> of the National flag and taking the state oath, followed by a two-minute <u>silence</u> to commemorate fallen War Heroes kicked off the day's sequence of events.
	Lexicon overlap	{'සහ': 'and'}
	Missed (Inflections)	නිශ්ශබ්දතාව ':'silence', 'එසවීම': 'hoist'

7 Conclusion

This research improved an existing multilingual embedding-based document alignment technique and a sentence alignment technique with the use of bilingual lexicons. The study was conducted focusing on the low-resource language pairs Sinhala–English, Sinhala–Tamil and Tamil–English. Since we experimented with LASER, XLM-R and LaBSE multilingual embeddings, our work serves as an empirical study on the effectiveness of these models for document and sentence alignment. Our results show that positive gains can be obtained even with bilingual lexicons having very small quantities of parallel phrases. We have also compiled and released gold-standard human annotated evaluation sets for document alignment and sentence alignment for the considered languages, which will enable future research in this context. As future work, we plan to further improve the multilingual representations and cross-lingual mappings of the PMLMs, for low-resource languages by exploring different fine-tuning objectives.

Acknowledgements Aloka Fernando was initially funded by the Accelerating Higher Education Expansion and Development (AHEAD) Operation of the Ministry of Education, Sri Lanka, funded by the World Bank. Currently, she is funded by a Senate Research Committee (SRC) grant from the University of Moratuwa, Sri Lanka. Dataset creation was funded by an SRC grant from University of Moratuwa, Sri Lanka.

Funding Funding was provided by Higher Education Expansion and Development (AHEAD) and Senate Research Committee (SRC) Grant University of Moratuwa.

Declarations

Conflict of interest The authors have no conflict of interest to declare.

Appendix A Algorithms for using bilingual lexicons

Our improvement to the document alignment and sentence alignment algorithms consider bilingual lexicons as explained in Sects. 4.1.1 and 4.2.2, respectively. The supporting algorithms related to term matching using person names (Algorithm 1) and rest of the bilingual lexicons (Algorithm 2) are shown below.

Algorithm 1 Calculate *count* for Equation 8 considering Person Names lexicon

Require: $s_A, s_B, dict$
$w_A \Leftarrow list(s_A)$
$w_B \Leftarrow list(s_B)$
$count \Leftarrow 0$
for $w \in w_A : |w| = 1$ **do**
  if $w \in dict$ **then**
    $v_B \Leftarrow dict[w]$
    for $v \in v_B$ **do**
      if $v \in w_B$ **then**
        $count \Leftarrow count + 1$
        Remove w from w_B
      end if
    end for
  end if
end for

Algorithm 2 Calculate *count* for Equation 8 considering Designations and Word Dictionary

Require: $s_A, s_B, dict$
$w_A \Leftarrow list(s_A)$
$w_B \Leftarrow list(s_B)$
$count \Leftarrow 0$
if $|w_A| \geq 5$ **then**
  for $w \in w_A : |w| = 1, 2, 3, 4, 5$ **do**
    if $w \in dict$ **then**
      $v_B \Leftarrow dict[w]$
      for $v \in v_B$ **do**
        if $v \in w_B$ **then**
          $count \Leftarrow count + 1$
          Remove w from w_B
        end if
      end for
    end if
  end for
else
  Algorithm 1
end if

Appendix B Document alignment results

Table 11 shows the document alignment results for each news source for the language pairs English–Sinhala, English–Tamil and Sinhala–Tamil. In Table 7, the individual scores obtained for the news sources are averaged. The score in bold is the result corresponding to the best $F1$ score with respective to the news source and language pair.

Table 11 Document Alignment results in terms of recall (R), precision (P) and $F1$ with respective to each language pair. Here, BL refers to the recreated Baseline [13] considering LASER embeddings. On top of this, each bilingual lexicon had been added and the experiments were repeated. The bilingual lexicons considered were Person Names (N), Designations (Ds), Dictionary (Dc) and Improved Dictionary (MDc). Subsequently considering PMLMs XLM-R and LaBSE, the same set of experiments have been conducted.

Experiment	Wt.	En–Si											
		Hiru			ITN			Newsfirst			Army		
		R	P	$F1$	R	P	$F1$	R	P	$F1$	R	P	$F1$
LASER													
BL	SL	82.25	71.06	76.24	91.22	37.28	52.93	96.01	47.37	63.44	**99.41**	**94.55**	**96.91**
	IDF	79.31	68.52	73.52	89.39	36.53	51.87	94.17	46.46	62.22	97.02	92.28	94.59
	SLIDF	82.32	71.12	76.31	91.22	37.28	52.93	95.89	47.31	63.36	**99.41**	**94.55**	**96.91**
BL+N	SL	84.90	73.35	78.70	92.78	37.92	53.84	96.31	47.52	63.64	99.19	94.34	96.70
	IDF	81.89	70.75	75.91	90.78	37.10	52.67	94.17	46.46	62.22	97.73	92.95	95.28
	SLIDF	84.90	73.35	78.70	92.87	37.95	53.88	96.31	47.52	63.64	99.19	94.34	96.70
BL+N+Ds	SL	84.90	73.35	78.70	92.78	37.92	53.84	96.31	47.52	63.64	99.19	94.34	96.70
	IDF	81.89	70.75	75.91	90.78	37.10	52.67	94.17	46.46	62.22	97.73	92.95	95.28
	SLIDF	84.90	73.35	78.70	92.87	37.95	53.88	96.31	47.52	63.64	99.19	94.34	96.70
BL+N+Ds+Dc	SL	85.61	73.96	79.36	93.13	38.06	54.04	96.55	47.64	63.80	99.35	94.49	96.86
	IDF	81.89	70.75	75.91	90.78	37.10	52.67	94.17	46.46	62.22	97.73	92.95	95.28
	SLIDF	84.90	73.35	78.70	92.87	37.95	53.88	96.31	47.52	63.64	99.19	94.34	96.70
BL+N+Ds+MDc	SL	**85.90**	**74.21**	**79.63**	**94.00**	**38.41**	**54.54**	**97.32**	**48.02**	**64.31**	99.35	94.49	96.86
	IDF	81.89	70.75	75.91	90.78	37.10	52.67	94.17	46.46	62.22	97.73	92.95	95.28
	SLIDF	84.90	73.35	78.70	92.87	37.95	53.88	96.31	47.52	63.64	99.19	94.34	96.70

Table 11 continued

Experiment	Wt.	En–Si											
		Hiru			ITN			Newsfirst			Army		
		R	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1
XLM-R													
BL	SL	91.05	78.66	84.41	98.09	40.09	56.91	98.39	48.55	65.01	99.46	94.60	96.97
	IDF	91.41	78.97	84.74	98.00	40.05	56.86	98.21	48.46	64.90	99.03	94.18	96.54
	SLIDF	90.91	78.54	84.27	98.09	40.09	56.91	98.39	48.55	65.01	99.46	94.60	96.97
BL+N	SL	92.77	80.15	86.00	**98.26**	**40.16**	**57.02**	98.57	48.63	65.13	**99.73**	**94.85**	**97.23**
	IDF	92.27	79.72	85.54	97.83	39.98	56.76	97.92	48.31	64.70	99.35	94.49	96.86
	SLIDF	92.91	80.27	86.13	**98.26**	**40.16**	**57.02**	98.57	48.63	65.13	**99.73**	**94.85**	**97.23**
BL+N+Ds	SL	92.77	80.15	86.00	**98.26**	**40.16**	**57.02**	98.57	48.63	65.13	**99.73**	**94.85**	**97.23**
	IDF	92.27	79.72	85.54	97.83	39.98	56.76	97.92	48.31	64.70	99.35	94.49	96.86
	SLIDF	92.91	80.27	86.13	**98.26**	**40.16**	**57.02**	98.57	48.63	65.13	**99.73**	**94.85**	**97.23**
BL+N+Ds+Dc	SL	92.63	80.03	85.87	98.26	40.16	57.01	98.51	48.60	65.09	**99.73**	**94.85**	**97.23**
	IDF	92.27	79.72	85.54	97.83	39.98	56.76	97.92	48.31	64.70	99.35	94.49	96.86
	SLIDF	92.91	80.27	86.13	**98.26**	**40.16**	**57.02**	98.57	48.63	65.13	**99.73**	**94.85**	**97.23**
BL+N+Ds+MDc	SL	**93.63**	**80.89**	**86.80**	98.17	40.12	56.96	**98.69**	**48.69**	**65.21**	**99.73**	**94.85**	**97.23**
	IDF	92.27	79.72	85.54	97.83	39.98	56.76	97.92	48.31	64.70	99.35	94.49	96.86
	SLIDF	92.91	80.27	86.13	**98.26**	**40.16**	**57.02**	98.57	48.63	65.13	**99.73**	**94.85**	**97.23**

Table 11 continued

Experiment	Wt.	En–Si											
		Hiru			ITN			Newsfirst			Army		
		R	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1
LaBSE													
BL	SL	95.42	82.44	88.45	98.78	40.37	57.32	**99.11**	**48.90**	**65.49**	**99.73**	**94.85**	**97.23**
	IDF	95.49	82.50	88.52	98.35	40.19	57.06	99.23	48.96	65.56	99.67	94.80	97.18
	SLIDF	95.35	82.38	88.39	98.78	40.37	57.32	**99.11**	**48.90**	**65.49**	**99.73**	**94.85**	**97.23**
BL+N	SL	95.42	82.44	88.46	98.87	40.41	57.37	98.99	48.84	65.41	**99.73**	**94.85**	**97.23**
	IDF	**95.71**	**82.68**	**88.72**	98.43	40.23	57.12	98.99	48.84	65.41	99.68	94.80	97.18
	SLIDF	95.42	82.44	88.46	98.87	40.41	57.37	98.99	48.84	65.41	**99.73**	**94.85**	**97.23**
BL+N+Ds	SL	95.42	82.44	88.46	98.87	40.41	57.37	98.99	48.84	65.41	**99.73**	**94.85**	**97.23**
	IDF	**95.71**	**82.68**	**88.72**	98.43	40.23	57.12	98.99	48.84	65.41	99.68	94.80	97.18
	SLIDF	95.42	82.44	88.46	98.87	40.41	57.37	98.99	48.84	65.41	**99.73**	**94.85**	**97.23**
BL+N+Ds+Dc	SL	95.28	82.31	88.32	98.96	40.44	57.42	**99.11**	**48.90**	**65.49**	**99.73**	**94.85**	**97.23**
	IDF	**95.71**	**82.68**	**88.72**	98.43	40.23	57.12	98.99	48.84	65.41	99.68	94.80	97.18
	SLIDF	95.42	82.44	88.46	98.87	40.41	57.37	98.99	48.84	65.41	**99.73**	**94.85**	**97.23**
BL+N+Ds+MDc	SL	95.49	82.50	88.52	**99.04**	**40.48**	**57.47**	98.99	48.84	65.41	**99.73**	**94.85**	**97.23**
	IDF	**95.71**	**82.68**	**88.72**	98.43	40.23	57.12	98.99	48.84	65.41	99.68	94.80	97.18
	SLIDF	95.42	82.44	88.46	98.87	40.41	57.37	98.99	48.84	65.41	**99.73**	**94.85**	**97.23**

Experiment	En–Ta											
	Hiru			ITN			Newsfirst			Army		
	R	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1
LASER												
BL	25.13	18.09	21.04	50.78	19.00	27.65	53.71	21.47	30.68	72.27	67.43	69.77
	22.65	16.31	18.96	52.62	19.68	28.65	52.45	20.97	29.96	64.51	60.20	62.28
	25.30	18.21	21.18	50.92	19.05	27.72	53.95	21.57	30.81	72.39	67.54	69.88

Table 11 continued

Experiment	En–Ta											
	Hiru			ITN			Newsfirst			Army		
	R	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1
BL+N	26.07	18.77	21.83	52.05	19.47	28.34	54.34	21.72	31.04	73.85	68.91	71.29
	24.70	17.78	20.68	54.03	20.21	29.42	52.76	21.09	30.13	64.81	60.47	62.56
	26.24	18.89	21.97	52.05	19.47	28.34	54.66	21.85	31.22	73.91	68.97	71.35
BL+N+Ds	26.07	18.77	21.83	52.05	19.47	28.34	54.34	21.72	31.04	73.85	68.91	71.29
	24.70	17.78	20.68	54.03	20.21	29.42	52.76	21.09	30.13	64.81	60.47	62.56
	26.24	18.89	21.97	52.05	19.47	28.34	54.66	21.85	31.22	73.91	68.97	71.35
BL+N+Ds+Dc	47.44	34.15	39.71	74.82	27.99	40.74	76.14	30.44	43.49	84.55	78.89	81.62
	44.36	31.94	37.14	74.26	27.78	40.43	72.20	28.86	41.24	77.97	72.75	75.27
	47.35	34.09	39.64	74.82	27.99	40.74	75.83	30.31	43.31	84.55	78.89	81.62
BL+N+Ds+MDc	50.85	36.62	42.58	**77.23**	**28.89**	**42.05**	80.25	32.08	45.84	**87.02**	**81.19**	**84.00**
	47.44	34.15	39.71	76.52	28.62	41.66	75.99	30.38	43.40	79.79	74.45	77.03
	50.94	**36.68**	**42.65**	**77.23**	**28.89**	**42.05**	**80.57**	**32.21**	**46.02**	**87.02**	**81.19**	**84.00**

Table 11 continued

Experiment	En–Ta											
	Hiru			ITN			Newsfirst			Army		
	R	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1
XLM-R												
BL	82.31	59.26	68.91	94.34	35.29	51.37	97.08	38.81	55.45	94.77	88.43	91.49
	81.62	58.77	68.34	95.33	35.66	51.91	96.92	38.74	55.36	95.36	88.98	92.06
	82.39	59.32	68.98	94.34	35.29	51.37	96.92	38.74	55.36	94.77	88.43	91.49
BL+N	82.82	59.63	69.34	94.63	35.40	51.53	97.08	38.81	55.45	95.53	89.14	92.22
	82.65	59.51	69.20	94.34	35.29	51.37	95.34	38.11	54.45	95.12	88.76	91.83
	82.91	59.69	69.41	94.63	35.40	51.53	96.92	38.74	55.35	95.53	89.14	92.22
BL+N+Ds	82.82	59.63	69.34	94.63	35.40	51.53	97.08	38.81	55.45	95.53	89.14	92.22
	82.65	59.51	69.20	94.34	35.29	51.37	95.34	38.11	54.45	95.12	88.76	91.83
	82.91	59.69	69.41	94.63	35.40	51.53	96.92	38.74	55.35	95.53	89.14	92.22
BL+N+Ds+Dc	85.04	61.23	71.20	97.31	36.40	52.98	97.71	39.06	55.81	97.42	90.90	94.04
	84.10	60.55	70.41	96.46	36.09	52.52	96.60	38.62	55.18	96.30	89.86	92.97
	85.04	61.23	71.20	97.31	36.40	52.98	**97.79**	**39.09**	**55.85**	97.42	90.90	94.04
BL+N+Ds+MDc	**85.21**	**61.35**	**71.34**	**97.45**	**36.45**	**53.06**	97.71	39.06	55.81	**97.53**	**91.01**	**94.16**
	83.93	60.43	70.27	96.89	36.24	52.75	96.68	38.65	55.22	96.18	89.75	92.85
	85.21	**61.35**	**71.34**	**97.45**	**36.45**	**53.06**	97.45	36.45	53.06	**97.53**	**91.01**	**94.16**

Table 11 continued

Experiment	En–Ta											
	Hiru			ITN			Newsfirst			Army		
	R	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1
LaBSE												
BL	**87.09**	**62.71**	**72.92**	99.58	37.25	54.22	98.10	39.22	56.03	98.47	91.89	95.07
	85.64	61.66	71.70	99.58	37.25	54.22	**98.10**	**39.22**	**56.03**	98.30	91.72	94.89
	87.01	62.65	72.84	98.10	**39.22**	**56.03**	**98.10**	**39.22**	**56.03**	**98.47**	91.89	95.07
BL+N	86.75	62.46	72.63	99.58	37.25	54.22	97.95	39.15	55.94	98.41	91.83	95.01
	85.81	61.78	71.84	99.15	37.09	53.99	96.68	38.65	55.22	98.18	91.61	94.78
	86.67	62.40	72.56	99.58	37.25	54.22	97.95	39.15	55.94	98.41	91.83	95.01
BL+N+Ds	86.75	62.46	72.63	99.58	37.25	54.22	97.95	39.15	55.94	98.41	91.83	95.01
	85.81	61.78	71.84	99.15	37.09	53.99	96.68	38.65	55.22	98.18	91.61	94.78
	86.67	62.40	72.56	99.58	37.25	54.22	97.95	39.15	55.94	98.41	91.83	95.01
BL+N+Ds+Dc	85.64	61.66	71.70	99.43	37.20	54.14	96.76	38.68	55.27	98.12	91.56	94.72
	86.41	62.22	72.34	**99.86**	37.35	54.37	97.95	39.15	55.94	98.41	91.83	95.01
	86.41	62.22	72.34	**99.86**	37.35	54.37	97.87	39.12	55.90	98.41	**95.01**	**96.68**
BL+N+Ds+MDc	86.41	62.22	72.34	99.86	37.35	54.37	97.87	39.12	55.90	98.41	**95.01**	**96.68**
	85.90	61.85	71.91	99.43	37.20	54.14	96.92	38.74	55.36	98.06	91.50	94.67
	86.15	62.03	72.13	**99.86**	37.35	54.37	97.87	39.12	55.90	98.41	91.83	95.01
Experiment	Si–Ta											
	Hiru			ITN			Newsfirst			Army		
	R	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1
LASER												
BL	43.71	32.61	37.35	84.68	30.12	44.44	82.82	28.24	42.12	82.45	72.24	77.01
	41.81	31.20	35.73	85.80	30.52	45.03	79.76	27.20	40.56	76.81	67.30	71.74
	43.81	32.69	37.44	84.68	30.12	44.44	82.03	27.97	41.72	82.45	72.24	77.01

Table 11 continued

Experiment	Si–Ta											
	Hiru			ITN			Newsfirst			Army		
	R	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1
BL+N	49.10	36.64	41.96	88.87	31.61	46.63	85.09	29.01	43.27	86.57	75.85	80.86
	46.40	34.63	39.66	90.40	32.16	47.44	83.42	28.44	42.42	80.86	70.85	75.52
	49.25	36.75	42.09	88.87	31.61	46.63	85.00	28.98	43.22	86.69	75.96	80.97
BL+N+Ds	49.10	36.64	41.96	88.87	31.61	46.63	85.09	29.01	43.27	86.57	75.85	80.86
	41.81	31.20	35.73	85.80	30.52	45.03	85.09	29.01	43.27	76.81	67.30	71.74
	49.25	36.75	42.09	88.87	31.61	46.63	85.00	28.98	43.22	86.69	75.96	80.97
BL+N+Ds+Dc	52.60	39.25	44.95	91.52	32.56	48.03	87.27	29.75	44.38	87.07	76.29	81.33
	50.35	37.57	43.03	92.44	32.88	48.51	85.19	29.05	43.32	82.13	71.96	76.71
	52.45	39.13	44.82	91.52	32.56	48.03	87.27	29.75	44.38	87.20	76.40	81.44
BL+N+Ds+MDc	**57.34**	**42.79**	**49.01**	93.97	33.43	49.32	**90.92**	**31.00**	**46.24**	**89.73**	**78.62**	**83.81**
	54.50	40.66	46.57	**94.38**	**33.58**	**49.53**	88.55	30.19	45.03	84.60	74.13	79.02
	57.24	42.71	48.92	93.97	33.43	49.32	**90.92**	**31.00**	**46.24**	89.67	78.57	83.75

Table 11 continued

Experiment	Si–Ta											
	Hiru			ITN			Newsfirst			Army		
	R	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1
XLM-R												
BL	78.77	58.78	67.32	98.47	35.03	51.68	98.81	33.69	50.25	92.65	81.18	86.53
	77.07	57.51	65.87	99.18	35.28	52.05	98.32	33.52	50.00	89.61	78.51	83.69
	78.82	58.81	67.36	98.47	35.03	51.68	98.81	33.69	50.25	92.65	81.18	86.53
BL+N	79.87	59.60	68.26	99.08	35.25	52.00	98.82	33.69	50.25	94.36	82.68	88.13
	78.87	58.85	67.40	99.18	35.28	52.05	98.32	33.52	50.00	91.63	80.29	85.59
	79.82	59.56	68.22	99.08	35.25	52.00	98.82	33.69	50.25	94.36	82.68	88.13
BL+N+Ds	79.87	59.60	68.26	99.08	35.25	52.00	98.82	33.69	50.25	94.36	82.68	88.13
	77.07	57.51	65.87	99.18	35.28	52.05	98.82	33.69	50.25	89.61	78.51	83.69
	79.82	59.56	68.22	99.08	35.25	52.00	98.82	33.69	50.25	94.36	82.68	88.13
BL+N+Ds+Dc	80.27	59.90	68.60	99.49	35.39	52.21	98.91	33.73	50.30	94.55	82.84	88.31
	78.82	58.81	67.36	99.18	35.28	52.05	98.42	33.56	50.05	92.08	80.68	86.00
	80.27	59.90	68.60	99.49	35.39	52.21	99.01	33.76	50.35	94.55	82.84	88.31
BL+N+Ds+MDc	**81.12**	**60.53**	**69.33**	**99.69**	**35.47**	**52.32**	**99.01**	**33.76**	**50.35**	**95.25**	**83.45**	**88.96**
	79.17	59.08	67.66	99.08	35.25	52.00	98.42	33.56	50.05	93.60	82.01	87.42
	81.02	60.45	69.24	**99.69**	**35.47**	**52.32**	**99.01**	**33.76**	**50.35**	**95.25**	**83.45**	**88.96**

Table 11 continued

Experiment	Si–Ta											
	Hiru			ITN			Newsfirst			Army		
	R	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1
LaBSE												
BL	87.36	65.19	74.66	99.50	35.57	52.41	99.41	33.89	50.55	**99.11**	**86.84**	**92.57**
	87.46	65.26	74.75	**99.97**	**35.60**	**52.50**	99.41	33.89	50.55	98.99	86.73	92.45
	87.36	65.19	74.66	**99.97**	**35.60**	**52.50**	99.41	33.89	50.55	**99.11**	**86.84**	**92.57**
BL+N	87.06	64.96	74.40	99.50	35.57	52.41	**99.51**	**33.93**	**50.61**	**99.11**	**86.84**	**92.57**
	87.36	65.19	74.66	99.18	35.28	52.05	99.41	33.89	50.55	98.48	86.29	91.98
	87.26	65.11	74.58	99.50	35.57	52.41	**99.51**	**33.93**	**50.61**	**99.11**	**86.84**	**92.57**
BL+N+Ds	87.06	64.96	74.40	99.50	35.57	52.41	**99.51**	**33.93**	**50.61**	**99.11**	**86.84**	**92.57**
	87.46	65.26	74.75	**99.97**	**35.60**	**52.50**	**99.51**	**33.93**	**50.61**	98.99	86.73	92.45
	87.26	65.11	74.58	99.50	35.57	52.41	**99.51**	**33.93**	**50.61**	**99.11**	**86.84**	**92.57**
BL+N+Ds+Dc	87.46	65.26	74.75	**99.97**	**35.60**	**52.50**	**99.51**	**33.93**	50.60	**99.11**	**86.84**	**92.57**
	87.56	65.34	74.83	99.90	35.54	52.43	**99.51**	**33.93**	50.60	98.48	86.28	91.98
	87.51	65.30	74.79	**99.97**	**35.60**	**52.50**	**99.51**	**33.93**	50.60	**99.11**	**86.84**	**92.57**
BL+N+Ds+MDc	**87.71**	**65.45**	**74.96**	**99.97**	**35.60**	**52.50**	**99.51**	**33.93**	50.60	**99.11**	**86.84**	**92.57**
	86.86	64.82	74.24	99.69	35.47	52.32	**99.51**	**33.93**	50.60	98.61	86.40	92.10
	87.66	65.41	74.92	**99.97**	**35.60**	52.50	99.31	33.86	50.50	**99.11**	**86.84**	**92.57**

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN et al (2017) Attention is all you need. Adv Neural Inf Process Syst 30
2. Koehn P, Knowles R (2017) Six challenges for neural machine translation. In: Proceedings of the first workshop on neural machine translation. Association for Computational Linguistics, Vancouver, pp 28–39
3. Ranathunga S, Lee ESA, Skenduli MP, Shekhar R, Alam M, Kaur R (2021) Neural machine translation for low-resource languages: a survey. arXiv preprint arXiv:2106.15115
4. Kreutzer J, Caswell I, Wang L, Wahab A, van Esch D, Ulzii-Orshikh N et al (2022) Quality at a glance: an audit of web-crawled multilingual datasets. Trans Assoc Comput Linguist 10:50–72
5. Bañón M, Chen P, Haddow B, Heafield K, Hoang H, Esplà-Gomis M et al (2020) ParaCrawl: web-scale acquisition of parallel corpora. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 4555–4567
6. Buck C, Koehn P (2016) Findings of the WMT 2016 bilingual document alignment shared task. In: Proceedings of the first conference on machine translation: volume 2, shared task papers. Association for Computational Linguistics, Berlin, pp 554–563
7. Resnik P (1998) Parallel strands: a preliminary investigation into mining the web for bilingual text. In: Conference of the association for machine translation in the Americas. Springer, pp 72–82
8. Resnik P (1999) Mining the web for bilingual text. In: Proceedings of the 37th annual meeting of the association for computational linguistics, pp 527–534
9. Papavassiliou V, Prokopidis P, Piperidis S (2016) The ilsp/arc submission to the wmt 2016 bilingual document alignment shared task. In: Proceedings of the first conference on machine translation: volume 2, shared task papers, pp 733–739
10. Resnik P, Smith NA (2003) The web as a parallel corpus. Comput Linguist 29(3):349–380
11. Espla-Gomis M, Forcada ML, Ortiz-Rojas S, Ferrández-Tordera J. (2016) Bitextor's participation in WMT'16: shared task on document alignment. In: Proceedings of the first conference on machine translation: volume 2, shared task papers, pp 685–691
12. Etchegoyhen T, Gete H (2020) Handle with care: a case study in comparable corpora exploitation for neural machine translation. In: Proceedings of The 12th language resources and evaluation conference, pp 3799–3807
13. El-Kishky A, Guzmán F (2020) Massively multilingual document alignment with cross-lingual sentence-mover's distance. In: Proceedings of the 1st conference of the asia-pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing. Association for Computational Linguistics, Suzhou, pp 616–625
14. Varga D, Halácsy P, Kornai A, Nagy V, Németh L, Trón V (2007) Parallel corpora for medium density languages. Amsterdam Stud Theory Hist Linguist Sci Ser 4(292):247
15. Munteanu DS, Marcu D (2005) Improving machine translation performance by exploiting non-parallel corpora. Comput Linguist 31(4):477–504
16. Sarikaya R, Maskey S, Zhang R, Jan EE, Wang D, Ramabhadran B et al (2009) Iterative sentence-pair extraction from quasi-parallel corpora for machine translation. In: Tenth annual conference of the international speech communication association, pp 432–435
17. Kvapilíková I, Artetxe M, Labaka G, Agirre E, Bojar O (2020) Unsupervised multilingual sentence embeddings for parallel corpus mining. In: Proceedings of the 58th annual meeting of the association for computational linguistics: student research workshop, pp 255–262
18. Feng F, Yang Y, Cer D, Arivazhagan N, Wang W (2020) Language-agnostic BERT sentence embedding. arXiv preprint arXiv:2007.01852
19. Artetxe M, Schwenk H (2019) Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Trans Assoc Comput Linguist 7:597–610
20. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F et al (2020) Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 8440–8451
21. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). Association for Computational Linguistics, Minneapolis, pp 4171–4186
22. Rajitha C, Piyarathna L, Sachintha D, Ranathunga S (2021) Metric learning in multilingual sentence similarity measurement for document alignment. In: Proceedings of the international conference on recent advances in natural language processing (RANLP 2021). Held online: INCOMA Ltd., pp 1150–1157. https://aclanthology.org/2021.ranlp-1.129

23. Ni J, Ábrego GH, Constant N, Ma J, Hall KB, Cer D et al (2021) Sentence-t5: scalable sentence encoders from pre-trained text-to-text models. arXiv preprint arXiv:2108.08877
24. Joshi P, Santy S, Budhiraja A, Bali K, Choudhury M (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, pp 6282–6293
25. Artetxe M, Schwenk H (2019) Margin-based parallel corpus mining with multilingual sentence embeddings. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Association for Computational Linguistics, pp 3197–3203
26. Koehn P, Khayrallah H, Heafield K, Forcada ML (2018) Findings of the wmt 2018 shared task on parallel corpus filtering. In: Proceedings of the third conference on machine translation: shared task papers, pp 726–739
27. Chen J, Nie JY (2000) Parallel web text mining for cross-language IR. In: Content-based multimedia information access, vol 1. RIAO, pp 62–77
28. Shi L, Niu C, Zhou M, Gao J (2006) A DOM tree alignment model for mining parallel data from the web. In: Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics, pp 489–496
29. Zafarian A, Sadeghi APA, Azadi F, Ghiasifard S, Panahloo ZA, Bakhshaei S et al (2015) AUT document alignment framework for BUCC workshop shared task. In: Proceedings of the eighth workshop on building and using comparable corpora, pp 79–87
30. Li B, Gaussier E (2013) Exploiting comparable corpora for lexicon extraction: Measuring and improving corpus quality. In: Building and using comparable corpora. Springer, pp 131–149
31. Ma X, Liberman M (1999) Bits: a method for bilingual text search over the web. In: Machine translation summit VII, pp 538–542
32. Fung P, Cheung P (2004) Mining very-non-parallel corpora: parallel sentence and lexicon extraction via bootstrapping and e. In: Proceedings of the 2004 conference on empirical methods in natural language processing, pp 57–63
33. Ion R, Ceauşu A, Irimia E (2011) An expectation maximization algorithm for textual unit alignment. In: Proceedings of the 4th workshop on building and using comparable corpora: comparable corpora and the web, pp 128–135
34. Gomes L, Lopes G (2016) First steps towards coverage-based document alignment. In: Proceedings of the first conference on machine translation: volume 2, shared task papers, pp 697–702
35. Morin E, Hazem A, Boudin F, Loginova-Clouet E (2015) LINA: identifying comparable documents from Wikipedia. In: Proceedings of the eighth workshop on building and using comparable corpora. Association for Computational Linguistics, Beijing, pp 88–91
36. Uszkoreit J, Ponte J, Popat A, Dubiner M (2010) Large scale parallel document mining for machine translation. In: Proceedings of the 23rd international conference on computational linguistics (Coling 2010), pp 1101–1109
37. Rajitha M, Piyarathna L, Nayanajith M, Surangika S (2020) Sinhala and English document alignment using statistical machine translation. In: 2020 20th international conference on advances in ICT for emerging regions (ICTer). IEEE, pp 29–34
38. Jakubina L, Langlais P (2016) Bad luc@ wmt 2016: a bilingual document alignment platform based on lucene. In: Proceedings of the first conference on machine translation: volume 2, shared task papers, pp 703–709
39. Medveď M, Jakubíček M, Kovář V (2016) English-French document alignment based on keywords and statistical translation. In: Proceedings of the first conference on machine translation: volume 2, shared task papers, pp 728–732
40. Buck C, Koehn P (2016) Quick and reliable document alignment via tf/idf-weighted cosine distance. In: Proceedings of the first conference on machine translation: volume 2, shared task papers, pp 672–678
41. Germann U (2016) Bilingual document alignment with latent semantic indexing. In: Proceedings of the first conference on machine translation: volume 2, shared task papers. Association for Computational Linguistics, Berlin, pp 692–696
42. Dara AA, Lin YC (2016) Yoda system for wmt16 shared task: bilingual document alignment. In: Proceedings of the first conference on machine translation: volume 2, shared task papers, pp 679–684
43. Brown PF, Lai JC, Mercer RL (1991) Aligning sentences in parallel corpora. In: 29th Annual meeting of the association for computational linguistics. Association for Computational Linguistics, Berkeley, pp 169–176
44. Gale WA, Church KW (1993) A program for aligning sentences in bilingual corpora. Comput Linguist 19(1):75–102

45. Ma X (2006) Champollion: a robust parallel text sentence aligner. In: Proceedings of the fifth international conference on language resources and evaluation (LREC'06). European Language Resources Association (ELRA), Genoa, pp 489–492
46. Munteanu DS, Marcu D (2002) Processing comparable corpora with bilingual suffix trees. In: Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP 2002), pp 289–295
47. Stefanescu D, Ion R, Hunsicker S (2012) Hybrid parallel sentence mining from comparable corpora. In: Proceedings of the 16th annual conference of the European association for machine translation, pp 137–144
48. Abdul-Rauf S, Schwenk H (2009) On the use of comparable corpora to improve SMT performance. In: Proceedings of the 12th conference of the european chapter of the ACL (EACL 2009), pp 16–23
49. Mahata S, Das D (2017) Bandyopadhyay S. Bucc2017: a hybrid approach for identifying parallel sentences in comparable corpora. In: Proceedings of the 10th workshop on building and using comparable corpora, pp 56–59
50. Azpeitia A, Etchegoyhen T, Garcia EM (2017) Weighted set-theoretic alignment of comparable sentences. In: Proceedings of the 10th workshop on building and using comparable corpora, pp 41–45
51. Azpeitia A, Etchegoyhen T, Garcia EM (2018) Extracting parallel sentences from comparable corpora with STACC variants. In: Proceedings of the 11th workshop on building and using comparable corpora, pp 48–52
52. Grégoire F, Langlais P (2017) Bucc 2017 shared task: a first attempt toward a deep learning framework for identifying parallel sentences in comparable corpora. In: Proceedings of the 10th workshop on building and using comparable corpora, pp 46–50
53. Iyyer M, Manjunatha V, Boyd-Graber J, Daumé III H (2015) Deep unordered composition rivals syntactic methods for text classification. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers), pp 1681–1691
54. Guoa M, Shenb Q, Yanga Y, Gea H, Cera D, Abregoa GH et al (2018) Effective parallel corpus mining using bilingual sentence embeddings. WMT 2018:165
55. Leong C, Wong DF, Chao LS (2018) Um-paligner: neural network-based parallel sentence identification model. In: 11th Workshop on building and using comparable corpora, p 53
56. Bouamor H, Sajjad H (2018) H2@ bucc18: parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In: Proceedings of workshop on building and using comparable corpora, pp 43–47
57. Hangya V, Fraser A (2019) Unsupervised parallel sentence extraction with parallel segment detection helps machine translation. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp 1224–1234
58. Schwenk H, Chaudhary V, Sun S, Gong H, Guzmán F (2021) 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, pp 1351–1361
59. Schwenk H, Wenzek G, Edunov S, Grave E, Joulin A, Fan A (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). Association for Computational Linguistics, pp 6490–6500
60. Yang Y, Ábrego GH, Yuan S, Guo M, Shen Q, Cer D et al (2019) Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In: Proceedings of the twenty-eighth international joint conference on artificial intelligence (IJCAI-19), pp 5370–5378
61. Zweigenbaum P, Sharoff S, Rapp R (2018) Overview of the third BUCC shared task: spotting parallel sentences in comparable corpora. In: Proceedings of 11th workshop on building and using comparable corpora, pp 39–42
62. Koehn P, Guzmán F, Chaudhary V, Pino J. (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), pp 54–72
63. Priyadarshani H, Rajapaksha M, Ranasinghe M, Sarveswaran K, Dias G (2019) Statistical machine learning for transliteration: transliterating names between Sinhala, Tamil and English. In: 2019 International conference on asian language processing (IALP). IEEE, pp 244–249
64. Farhath F, Ranathunga S, Jayasena S, Dias G (2018) Integration of bilingual lists for domain-specific statistical machine translation for Sinhala-Tamil. In moratuwa engineering research conference (MERCon). IEEE, pp. 538–543
65. Thompson B, Koehn P (2019) Vecalign: improved sentence alignment in linear time and space. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp 1342–1348

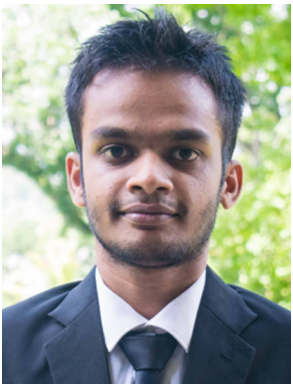
66. Fernando A, Ranathunga S, Dias G (2020) Data augmentation and terminology integration for domain-specific Sinhala-English-Tamil statistical machine translation. arXiv preprint arXiv:2011.02821
67. Guzmán F, Chen PJ, Ott M, Pino J, Lample G, Koehn P et al (2019) The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, pp 6098–6111
68. Goyal N, Gao C, Chaudhary V, Chen PJ, Wenzek G, Ju D et al (2021) The flores-101 evaluation benchmark for low-resource and multilingual machine translation. arXiv preprint arXiv:2106.03193
69. Liu Y, Gu J, Goyal N, Li X, Edunov S, Ghazvininejad M et al (2020) Multilingual denoising pre-training for neural machine translation. Trans Assoc Comput Linguist 8:726–742
70. Thillainathan S, Ranathunga S, Jayasena S (2021) Fine-tuning self-supervised multilingual sequence-to-sequence models for extremely low-resource NMT. In: 2021 Moratuwa engineering research conference (MERCon). IEEE, pp 432–437
71. Lee ESA, Thillainathan S, Nayak S, Ranathunga S, Adelani DI, Su R et al (2022) Pre-trained multilingual sequence-to-sequence models: a hope for low-resource language translation? arXiv preprint https://doi.org/10.48550/arXiv.2203.08850
72. Ott M, Edunov S, Baevski A, Fan A, Gross S, Ng N et al (2019) fairseq: a fast, extensible toolkit for sequence modeling. In: Proceedings of NAACL-HLT 2019: demonstrations, pp 48–53
73. Post M (2018) A call for clarity in reporting BLEU scores. In: Proceedings of the third conference on machine translation: research papers. Association for Computational Linguistics, Belgium, pp 186–191

Aloka Fernando obtained her BSc in Engineering (Hons) specializing in Electrical Engineering from the University of Moratuwa, Sri Lanka. She is currently a PhD candidate at the same University. Her current research is on low-resource machine translation and had contributed to linguistic resource development related to local languages in Sri Lanka. Prior to her tenure in the research domain, she had been working in the software industry for 8 years in diverse capacities.

Surangika Ranathunga received her BSc in Engineering (Hons) majoring in Computer Science and Engineering, and MSc in Computer Science from University of Moratuwa, Sri Lanka. She received her PhD from University of Otago, New Zealand. She is currently a senior lecturer at the Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka. Her research interests include Natural Language Processing, Information Retrieval, and Machine Learning.

Dilan Sachintha obtained his BSc in Engineering (Hons) majoring in Computer Science and Engineering from University of Moratuwa, Sri Lanka. Currently, he is working as a Software Engineer in WSO2 Lanka (Pvt) Ltd.

Lakmali Piyarathna obtained her BSc in Engineering (Hons) majoring in Computer Science and Engineering from University of Moratuwa, Sri Lanka. Currently, she is working as a Software Engineer in WSO2 Lanka (Pvt) Ltd.



Charith Rajitha obtained his BSc in Engineering (Hons) majoring in Computer Science and Engineering from University of Moratuwa, Sri Lanka. Currently, he is working as a Software Engineer in WSO2 Lanka (Pvt) Ltd.