Sinhala-English Word Embedding Alignment: Introducing Datasets and Benchmark for a Low Resource Language

Kasun Wickramasinghe and **Nisansa de Silva**
Department of Computer Science and Engineering
University of Moratuwa
Sri Lanka
{kasunw.22,NisansaDdS}@cse.mrt.ac.lk

Abstract

Since their inception, embeddings have become a primary ingredient in many flavours of Natural Language Processing (NLP) tasks supplanting earlier types of representation. Even though multilingual embeddings have been used for the increasing number of multilingual tasks, due to the scarcity of parallel training data, low-resource languages such as Sinhala, tend to focus more on monolingual embeddings. Then when it comes to the aforementioned multilingual tasks, it is challenging to utilize these monolingual embeddings given that even if the embedding spaces have a similar geometric arrangement due to an identical training process, the embeddings of the languages considered are not aligned. This is solved by the embedding alignment task. Even in this, high-resource language pairs are in the limelight while low-resource languages such as Sinhala which is in dire need of help seem to have fallen by the wayside. In this paper, we try to align Sinhala and English word embedding spaces based on available alignment techniques and introduce a benchmark for Sinhala language embedding alignment. In addition to that, to facilitate the supervised alignment, as an intermediate task, we also introduce Sinhala-English alignment datasets. These datasets serve as our anchor datasets for supervised word embedding alignment. Even though we do not obtain results comparable to the high-resource languages such as French, German, or Chinese, we believe our work lays the groundwork for more specialized alignment between English and Sinhala embeddings.

1 Introduction

Embedding spaces have been shown to have similar geometric arrangements (Mikolov et al., 2013b; Lample et al., 2018) especially when the training process is similar but, separately trained spaces are not aligned by default and that is a huge burden when it comes to certain multilingual tasks where having aligned embeddings are required.

Aligned embeddings are useful in multilingual tasks since similar words and sentences in each language can be considered to reside closer to each other in a common embedding space. So that we can do mathematical operations on the embeddings regardless of the language (Feng et al., 2022; Conneau and Lample, 2019).

The alignment is required for two types of embedding models:

1. Embedding models separately trained on monolingual data (Mikolov et al., 2013a; Bojanowski et al., 2017) and
2. Multilingual embedding models trained on parallel multilingual data (Feng et al., 2022; Conneau and Lample, 2019; Conneau et al., 2020).

As far as the multilingual models are concerned, most of the time the training process itself implicitly encourages alignment (Feng et al., 2022; Conneau and Lample, 2019). Conversely, when the monolingual models are concerned, the alignment has to be done explicitly after the models are trained. Multilingual models (Feng et al., 2022; Conneau and Lample, 2019; Conneau et al., 2020) are becoming more common for multilingual tasks nowadays due to the aforementioned implicit alignment of the training process (Feng et al., 2022; Conneau and Lample, 2019).

Monolingual embedding models have been there for decades and aligning monolingual embedding models is beneficial in various aspects rather than using multilingual models.

- Monolingual models are lightweight
- Can be run using simpler libraries and frameworks
- Using multilingual models may be redundant due to supporting many languages (Feng et al.,

2022; Conneau and Lample, 2019; Conneau et al., 2020)

- Multilingual model accuracy can be compromised due to the support of many languages (Feng et al., 2022)

- The accuracy for low-resource languages can be less compared to high-resource languages due to training data imbalance (Feng et al., 2022) in multilingual models (Eg: ∼700 Sinhala tokens in XLM-R (Conneau et al., 2020) vocabulary)

- Training or fine-tuning a multilingual model is time and resource-consuming (Feng et al., 2022; Conneau and Lample, 2019; Conneau et al., 2020)

Therefore, aligning existing monolingual models is still vital. *Aligned* word embedding models for common high-resource languages are officially provided by FastText[1] but most of the aligned low-resource language models are not publicly available. *Sinhala* being such a low-resource language, suffers from the aforementioned difficulties (de Silva, 2019; Ranathunga and de Silva, 2022). Several related works to the Sinhala language have been done previously by Smith et al. (2016) using *Procrusts* and Liyanage et al. (2021) using *VecMap* but, our attempt to properly make everything ready and available for future research. Therefore, our effort here is to,

- Set a benchmark for Sinhala word embedding alignment

- Introduce dataset induction methods for low-resource languages when parallel word corpora are not available

- Introduce MUSE[2]-like (Lample et al., 2018) alignments datasets for Sinhala-English language pair

- Provide aligned embeddings for Sinhala-English pair

- Release the code-base[3] related to all the experiments we have conducted.

This is more so the case for low-resource languages such as Sinhala (de Silva, 2019). This problem gets further accentuated due to the unreliable nature of the quality of existing parallel corpora for such low-resource languages (Kreutzer et al., 2022).

2 Related Work

2.1 Embedding Generation

The first major turning point in the word embedding domain was the introduction of Word2Vec by Mikolov et al. (2013a).Subsequently, two new Word2Vec-like embedding models were released which are the well-known Glove (Pennington et al., 2014) and FastText (Bojanowski et al., 2017) models. Those are global embedding models.

The idea behind *Embeddings from Language Models (ELMo)* (Peters et al., 2018) is generating a *context-based embedding* for a given word. In the transformers Vaswani et al. (2017) era, the first member of context-based transformer encoders is the *BERT* (Devlin et al., 2019) which is a stack of transformer encoders trained on two objectives named *Masked Language Modeling (MLM)* and *Next Sentence Prediction*. After that many variants of *BERT* have been released including *sentence transformers* (Reimers and Gurevych, 2019).

2.2 Word Embedding Alignment Techniques

For word embedding alignment, there have been different approaches since the release of Word2Vec (Mikolov et al., 2013a). The first work we come across is the work by Mikolov et al. (2013b) in 2013. In the following subsections, we are talking about the major approaches that have been there for word embedding alignment.

2.2.1 Simple Linear Mapping

Our first method is to find a linear mapping W, assuming the geometric arrangements of two embedding spaces are similar as per Mikolov et al. (2013b). The optimizing objective, therefore, is to minimize the Euclidean distance between the target and the mapped vectors as per Equation 1.

$$\underset{W}{min} \sum_{i=1}^{n} \|Wx_i - z_i\|^2 \tag{1}$$

2.2.2 Orthogonal Mapping

The second method we are trying is, finding an *orthogonal* mapping between the *normalized* source

[1] https://bit.ly/3LGqDrE
[2] https://github.com/facebookresearch/MUSE
[3] https://bit.ly/3t3SKu7

and the target embedding spaces (Xing et al., 2015). The major improvement we can expect from this mapping is that the optimizing objective is, from one perspective, optimizing the cosine distance between the target and the mapped embedding. The optimizing objective is as per Equation 2.

$$\underset{W}{max} \sum_{i} (Wx_i)^T z_i \tag{2}$$

2.2.3 Orthogonal Procrustes Mapping

In this case, the orthogonal transformation matrix is approximated using the product UV^T, where U and V are the transformation matrices of singular value decomposition (SVD) of the product X^TY where X and Y are the original source and target embeddings (Smith et al., 2016). As we know the U and V^T matrices only perform translation, rotation, uniform scaling, or a combination of these transformations, and no deformations are performed. Therefore the UV^T will simply align one embedding space to the other with the assumption that the geometric arrangement of the two spaces is similar.

2.2.4 CSLS Optimization

The third method we are trying is minimizing the *Cross-domain similarity local scaling* (CSLS) loss (Equation 3) as the optimization criterion (Joulin et al., 2018a). The mapping is assumed to be *orthogonal* and the emending is assumed to be *normalized*.

$$\begin{aligned} &\underset{W \in O_d}{min} \frac{1}{n} \sum_{i=1}^{n} -2x_i^T W^T y_i + \\ &\frac{1}{k} \sum_{y_j \in N_Y(Wx_i)} x_i^T W^T y_j + \\ &\frac{1}{k} \sum_{Wx_j \in N_X(y_i)} x_j^T W^T y_i \end{aligned} \tag{3}$$

Joulin et al. (2018a) have addressed the so-called *hubness problem* in embedding alignment. *Hubs* are words that appear too frequently in the neighbourhoods of other words. There have been solutions to mitigate this issue at *inference* by using different criteria (loss) such as Inverted Softmax (IFS) or CSLS, rather than using the same criteria used at the training phase. Using different criteria for inference adds an inconsistency. Therefore Joulin et al. (2018a) have included the CSLS criteria directly to the training objective and have achieved better results compared to previous related work. This is one of the alignment techniques used by *FastText* for their official *aligned word vectors*.

2.2.5 Unsupervised Techniques

The fourth method we are trying is the unsupervised alignment method where a parallel dictionary is not needed for the alignment where creating a quality parallel dictionary may consume extra time and resources. Unsupervised alignment can be done using,

- **Traditional statistical optimization techniques:** Artetxe et al. (2018) use an unsupervised initialization for the seed words based on the word similarity distributions claiming that the similar words of two languages should have similar distributions and then improve the mapping in an iterative manner using a self-learning technique. This method has been published as a framework called *VecMap*[4].

  The work by Grave et al. (2019) is about Procrustes analysis which learns a linear transformation between two sets of matched points $X \in R^{nXd}$ and $Y \in R^{nXd}$. If the correspondences between the two sets are known (i.e., which point of X corresponds to which point of Y), then the linear transformation can be recovered using least square minimization or finding the orthogonal mapping between the two spaces just like in supervised methods described just above. In this case, we do not know the correspondence between the two sets, nor the linear transformation. Therefore, the goal is to learn an orthogonal matrix $Q \in O_d$, such that the set of points X is close to the set of points Y and 1-to-1 correspondences (permutation matrix) can be found. They use the *Wasserstein distance* or *Earth Mover Distance* as the measure of distance between our two sets of points and then combine it with the orthogonal Procrustes, leading to the problem of Procrustes in Wasserstein distance or Wasserstein Procrustes (WP).

  Aboagye et al. (2022) have proposed Quantized Wasserstein Procrustes (qWP) Alignment which reduces the computational cost of the permutation matrix approximation in WP by quantizing the source and target embedding spaces.

[4] https://github.com/artetxem/vecmap

- **Adversarial methods:** One of the well-known unsupervised techniques is adversarial techniques where a *Generator* tries to mimic the desired results while a *Discriminator* tries to distinguish the real results from the generator results. The contest between the *Generator* and the *Discriminator* ends up having a *Generator* that can generate almost similar real results which the *Discriminator* can no longer distinguish. The work by Lample et al. (2018) follows an adversarial approach where they have obtained similar accuracy numbers as supervised alignment techniques by then.

2.3 English-Sinhala Embedding Alignment

Smith et al. (2016) have published[5] EN-Si alignment matrix along with 77 other languages. However, they have only worked in the Si→En direction (i.e. mapping En as the target). Their alignment datasets have not been published and most of the later experiments have been done using the *MUSE* datasets. Both MUSE and Smith et al. (2016) not having published an En-Si dataset we have to create our own dataset for supervised alignments as well as alignment result evaluation. Recently Liyanage et al. (2021) have experimented VecMap to align English and Sinhala embedding spaces for lexicon induction task

2.4 Alignment Datasets

The works by Guzmán et al. (2019), Hameed et al. (2016), Bañón et al. (2020) and Vasantharajan et al. (2022) that are comprised of sentence and paragraph level parallel entries. Apart from that there are several sentence and document-level parallel corpora available in OPUS[6]. They are well suited for higher-level multilingual tasks like Machine Translation (MT) but, not for lower-level tasks like word embedding alignment.

When it comes to word-level parallel corpora or simply dictionaries, we can find very few open-source resources for English-Sinhala language pairing. For most of the common language pairs, common alignment datasets have been published by MUSE but Sinhala is not available there. The dictionary *Subasa Ingiya*[7] (Wasala and Weerasinghe, 2008) is one of them which is a small dictionary that contains about 36000 pairs and contains not only word pairs but also phrases. The next resource is by Wickramasinghe and De Silva (2023) which introduces several pure word-level dictionaries.

3 Methodology

In this section, we present the methodologies we followed to obtain,

1. An alignment dataset for supervised embedding alignment
2. The alignment matrix between English and Sinhala word embedding spaces

Our primary research objective is to have an aligned Sinhala word embedding space with another high-resource language word embedding space such as English. We are experimenting with some of the techniques mentioned in Section 2.2. For the supervised techniques we need a parallel word corpus where each parallel pair acts as so-called *Anchor words*. For that purpose, we are creating an English-Sinhala parallel word dictionary which is our first task. The results we obtained and comparison with existing results are presented in Section 4.

3.1 Alignment Dataset Creation

Our first task is to create an alignment dataset for the supervised alignment. We experimented with two statistical methods and one available dataset adaptation to form the parallel word dictionary alias, our *alignment dataset*. In this section, we are presenting those techniques.

3.1.1 Pointwise Mutual Information Criterion

Pointwise Mutual Information (PMI) is used to identify how given two events are associated with each other. In Natural Language Processing (NLP) this measure is slightly improved as positive PMI where negative PMI values are clipped to 0 and this measure is used to identify context words of a given word.

$$\begin{aligned} pmi(x,y) &= log_2\left(\frac{P(x,y)}{P(x)P(y)}\right) \\ &= log_2\left(\frac{N.count(x,y)}{count(x).count(y)}\right) \end{aligned} \tag{4}$$

$$\begin{aligned} &ppmi(src,tgt) = max\left\{pmi(src,tgt),0\right\} \\ &= max\left\{log_2\left(\frac{N.count(src,tgt)}{count(src).count(tgt)}\right),0\right\} \end{aligned} \tag{5}$$

[5] https://bit.ly/3PTRW3Y

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

[7] https://subasa.lk/?page_id=3738

We used the PPMI measure between source and target word pairs in several parallel English-Sinhala corpora and by applying a threshold to PPMI we tried to obtain the corresponding translation (i.e. target word) for each source word.

Even if there are many sentence and paragraph-level parallel corpora out there, by considering the *size* and *quality (alignment)*, we selected only the following English-Sinhala parallel corpora to extract the dictionaries.

1. CCAligned-v1[8] - by El-Kishky et al. (2020)

2. OpenSubtitles-v2018[9] - Initially by Tiedemann (2016)

In our case, the N should be the total number of data points in the parallel corpus. Hence it becomes a global context rather than a local context. We observed that the dictionary building becomes unstable, i.e. many false pairs along with few correct pairs in the result. Therefore, we experimented with another approach that pays more attention to the local context.

3.1.2 Conditional Probability Product

In this approach, we have made a simple but valid assumption. That is, *"In a parallel corpus, the corresponding word translation pairs should co-occur"*. In other words, *"If two source and target language words co-occur more often, then there is a high chance for them to be a translation pair"*. If we can have a large enough corpus then we can say that this measurement tends to be more accurate due to the sampling statistics being closer to population statistics. Based on this assumption, we can find word translation pairs, as utilized in the corresponding optimization criterion in Equation 6, by finding the source-target word pairs that maximize the product of the two conditional probabilities:

1. Finding the target word in the context of the source word (corresponding translation) given the source word - $P(target|source)$

2. Finding the source word in the context of the target word (corresponding translation) given the target word - $P(source|target)$

$$\begin{aligned}
&\max_{src,tgt} \left[P\left(src|tgt\right) P\left(tgt|src\right)\right] \\
&\implies \max_{src,tgt} \left[\frac{P(src,tgt)^2}{P(source)P(target)}\right] \\
&\implies \max_{src,tgt} \left[\frac{count(src,tgt)^2}{count(src).count(tgt)}\right]
\end{aligned} \tag{6}$$

We used the same two corpora, *CCAligned* and *OpenSubtitles*, used in *ppmi* method explained in Section 3.1.1 to build the dictionaries here as well. This dataset is referred to as *Prob-based-dict* throughout the paper.

3.1.3 Using an Available Dataset

Recent work by Wickramasinghe and De Silva (2023) has introduced three English-Sinhala parallel dictionary datasets and the FastText version of that can be used for our work directly. They have published the datasets in GitHub[10].

Subsets of their dataset have been used to perform the embedding alignment. When building the alignment dataset we used 5k unique source words in the trainset and 1.5k unique source words in the test set. Not only that in the training set, we built the dataset purposefully including the most frequent English and Sinhala words. That is how MUSE datasets have been built as well. The datasets derived from this have been referred to with *En-Si-para* and *Si-En-para* prefixes in the paper.

3.1.4 Dataset Statistics

The statistics of the dataset are shown in Table 1. We have shown the unique word percentage with and without stop-words and, the lookup-precision with respect to the FastText (Bojanowski et al., 2017; Joulin et al., 2017) vocabularies as described in Equation 7. Spacy[11](En) and work by Lakmal et al. (2020) (Si) have been used for stop-word removal wherever necessary.

The *Look-up Precision*, P_L means, the proportion of *a word present in the FastText vocabulary, given that word is present in our alignment dictionary*. It is explained in Equation 7. The same thing can be simplified according to Equation 8 where N_{vocab} is the alignment dataset vocabulary size and $N_{available}$ is the number of dataset vocabulary words available in FastText vocabulary.

[8] https://bit.ly/3PECd7P
[9] https://bit.ly/3PEE5xh
[10] https://bit.ly/3ZDPrX2
[11] https://spacy.io/models/en

$$P_L = P\Big(\frac{\text{word present in the FastText vocabulary}}{\text{word present in the dictionary}}\Big) \tag{7}$$

$$P_L = coverage = \frac{N_{available}}{N_{vocab}} \tag{8}$$

3.2 Embedding Alignment

We have conducted the embedding alignment with FastText embeddings for English (En) (cc[12], wiki[13]) and Sinhala (Si) (cc[14], wiki[15]) trained on *Common Crawl*[16] (cc) and *Wikipedia*[17] (wiki) with the same setups followed by Joulin et al. (2018a).

- Learning rate in {1, 10, 25, 50} and number of epochs in {10, 20}
- Center the word vectors (optional)
- The number of nearest neighbours in the CSLS loss is 10
- Use the l2-normalized word vectors
- Use 200k word vectors for the training

We adopted our scripts from the alignment scripts by MUSE and FastText[18]. One major observation was that when we use an alignment dataset that consists of the most common words in languages, we obtain a higher test accuracy than having an alignment dataset without considering the most frequent words in languages.

4 Experiments

In this section, we present the experiments we have conducted and the obtained results and observations. We are using the FastText official embeddings of (Bojanowski et al., 2017). FastText provides two main embedding models: 1) Embeddings trained on Wikipedia (*wiki*), 2) Embeddings trained on Common-Crawl (*cc*).

Most of the previous related work has been done using the wiki embeddings but, when it comes to Sinhala wiki FastText embeddings, there are only 79030 word vectors in the official model (this is because the Sinhala content on Wikipedia is very low: To get an idea, the number of English articles at the moment are more than 6.5M while the number of Sinhala articles are just around 20k) but, the cc Sinhala model contains 808044 word vectors and therefore the wiki vectors are not rich enough for Sinhala. The experimental results also prove that fact. Due to that fact, in some comparisons, we are presenting the results obtained from the cc model.

Sinhala is morphologically richer than English and therefore the alignment is comparatively difficult. In most cases, a single English word can have multiple Sinhala representations. In that case, it is *not a good measure* to check the @1 precision on the test set to evaluate the alignment quality. Therefore checking a higher top-k precision (like @5 or @10) will be a better measure. The Procrustes alignment evaluation by Smith et al. (2016) also shows comparatively low @1 precision for Sinhala (language code Si - recall that they have performed the Si→En mapping). According to Aboagye et al. (2022) results, work by Joulin et al. (2018a) gives the best alignment results and therefore we have used Joulin et al. (2018a) as the main reference paper for our work here.

4.1 Dataset Comparison

As explained in section 3.1, we have created the alignment datasets in 3 different approaches, PPM based, conditional probability-based, and using a subset of the dataset by Wickramasinghe and De Silva (2023). In the first experiment, we evaluated all the datasets by aligning the English and Sinhala embeddings using the Procrustes (see section 2.2.3) method. The results are shown in Table 2.

We can see that the best accuracies have been shown by the *En-Si-para-cc-5k* and *En-Si-para-wiki-5k* datasets and therefore, for the rest of the experiments we have used the datasets created using Wickramasinghe and De Silva (2023) dataset.

4.2 Alignment Results

Table 3 reports the look-up/translation precision of the aligned wiki and cc English-Sinhala embeddings with different alignment techniques and retrieval criteria. The term after the last plus sign is the retrieval criteria. We can see that cc vectors show better alignment than wiki vectors. Table 4 shows the translation precision of different alignment techniques. RCSLS gives the best alignment

[12] https://bit.ly/3PCZ6It
[13] https://bit.ly/46aJzGX
[14] https://bit.ly/2JXAyL8
[15] https://bit.ly/48BGFgf
[16] https://commoncrawl.org/
[17] https://www.wikipedia.org/
[18] https://bit.ly/3Zz21Xe

Dataset	Language	Entries		Unique% w.r.t. stopwords		$P_L\%$ (Each language)		$P_L\%$ (Both languages)	
		Unique	Total	With	Without	wiki*	cc‡	wiki*	cc‡
Prob-based-dict	English	36713	67404	54.47	54.78	99.99	99.99	47.70	99.99
	Sinhala	53612	67404	79.54	79.67	39.22	100		
en-si-para-cc-5k†	English	5000	12803	39.05	39.67	100.00	100.00	100.00	100.00
	Sinhala	11403	12803	89.07	89.21	100.00	100.00		
en-si-para-wiki-5k†	English	5000	12782	39.12	39.74	100.00	100.00	100.00	100.00
	Sinhala	11394	12782	89.14	89.28	100.00	100.00		
si-en-para-cc-5k†	English	2406	6113	39.36	40.73	100.00	100.00	100.00	100.00
	Sinhala	5000	6113	81.81	81.96	100.00	100.00		
si-en-para-wiki-5k†	English	2397	6104	39.27	40.63	100.00	100.00	100.00	100.00
	Sinhala	5000	6104	81.93	82.09	100.00	100.00		

Table 1: Dataset Statistics: Statistics of the alignment datasets we have experimented with
* w.r.t. wiki-based FastText vocabulary ‡ w.r.t. common-cawl FastText vocabulary
† Subsets of Wickramasinghe and De Silva (2023)

Dataset	Retrieval	
	NN	CSLS
Prob-based-dict	13.6	16.7
En-Si-para-cc-5k	**16.4**	**20.4**

Table 2: En-Si Procrustes Embedding Alignment Results of cc-Fasttext embeddings on different datasets

in En→Si direction while the refined Procrustes method gives the best accuracy in Si→En direction. Table 5 shows a comparison between the Si-En alignment performed by Smith et al. (2016). They have reported the alignment results in Si→En direction only and also provided the alignment matrix associated with the alignment. The evaluation done using that alignment matrix and our evaluation dataset (rows 2, 3 of Table 5) may not reflect the exact accuracy since the original alignment dataset used by Smith et al. (2016) is not published and, therefore we cannot guarantee that our evaluation set and their training set are disjoint. Table 7 in Appendix A has further relevant analysis. Figure 1 shows the top-k retrieval distribution in both source-target and target-source directions of the aligned embeddings on the test sets for RCSL+NN and RCSL+CSLS using cc-FastText embeddings.

4.3 Impact of Alignment Dataset Size

In this section, we experimented with how the alignment dataset affects the alignment. We have experimented with an extended alignment dataset and evaluated it with the same test sets used in Section 4.2. The results are reported in Table 6.

5 Discussion and Future Work

According to Table 3 and 4, we observe that Si-En alignment results are not on par with the high-resource language pairs. We have identified several possible reasons for this score difference.

5.1 Impact of the embedding model size

We observe cc Fasttext models have better alignment than wiki models. According to Table 3 results we can see 22.6% @1 reduction (22.6→17.5) in En-Si direction and 41.5% @1 reduction (28.9→16.9) in En-Si direction. This effect can be expected due to the comparatively low (9.7% of cc vocabulary) vocabulary size of the Sinhala wiki FastText model (wiki-79k, cc-808k) and therefore missing a great portion of information on the Si side.

5.2 Quality of the alignment dataset

We have experimented only with the supervised alignment techniques in this paper and, the final alignment output solely depends on the quality of the alignment datasets that are used. Our main alignment experiments have been carried out using alignment datasets created using the base datasets provided by Wickramasinghe and De Silva (2023) and, according to their paper, it is mentioned that the so-called *look-up score* of the datasets are not higher as expected. That indicates that there is an issue with the quality/coverage of the base dataset we used. According to Smith et al. (2016) the more common word pairs in the alignment dataset the better the alignment output we achieve. How we

Method	wiki						cc					
	En-Si			Si-En			En-Si			Si-En		
	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10
Procrustes + NN	11.4	26.4	33.2	12.5	29.6	37.1	16.4	35.7	43.6	21.3	39.9	47.4
Procrustes + CSLS	14.8	31.5	39.8	14.4	27.6	33.8	20.4	39.9	**49.1**	18.0	31.9	37.4
Procrustes+ refine + NN	13.7	25.5	31.3	15.8	33.0	39.3	19.3	34.9	42.3	**28.9**	**45.7**	51.3
Procrustes+ refine + CSLS	16.1	29.0	35.7	**16.9**	31.0	36.7	20.9	38.6	46.3	21.7	36.6	41.6
RCSLS + spectral + NN	14.8	29.7	36.8	13.3	33.7	42.8	21.4	40.2	48.5	23.3	44.8	52.7
RCSLS + spectral + CSLS	17.1	33.1	41.0	15.1	29.4	35.1	21.5	41.7	49.1	19.2	34.9	41.8
RCSLS + NN	15.3	30.4	37.5	13.2	**34.1**	**43.3**	21.5	40.9	48.3	23.3	44.9	**53.2**
RCSLS + CSLS	**17.5**	**33.4**	**41.3**	15.5	29.3	35.9	**22.6**	**42.3**	**49.1**	19.4	35.4	42.1

Table 3: **English-Sinhala word translation average precisions** (@1, @5, @10) from 1.5k source word queries using 200k target words in **wiki** and **cc** Fasttext embeddings. *Refine* is the refinement step of Lampele et al. (2018) and, *Spectral* is the *Convex relaxation* step explained in Joulin et al. (2018b). For supervised alignments, two different train-test dataset pairs have been used.

Method	Joulin et al. (2018a)										Ours	
	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-zh	zh-en	en-si	si-en
Adv.+refine	81.7	83.3	82.3	82.1	74.0	72.2	44.0	59.1	32.5	31.4	-	-
Wass. Proc.+refine	82.8	84.1	82.6	82.9	75.4	73.3	43.7	59.1	-	-	-	-
Procrustes	81.4	82.9	81.1	82.4	73.5	72.4	51.7	63.7	42.7	36.7	20.4	18.0
Procrustes+ refine	82.4	83.9	82.3	83.2	75.3	73.2	50.1	63.5	40.3	35.5	20.9	**21.7**
RCSLS + spectral	83.5	85.7	82.3	**84.1**	78.2	75.8	56.1	66.5	44.9	45.7	21.5	19.2
RCSLS	**84.1**	**86.3**	**83.3**	**84.1**	**79.1**	**76.3**	**57.9**	**67.2**	**45.9**	**46.4**	**22.6**	19.4

Table 4: Extended Comparison among different alignment techniques using CSLS retrieval. Here only the top-1 precision scores have been included

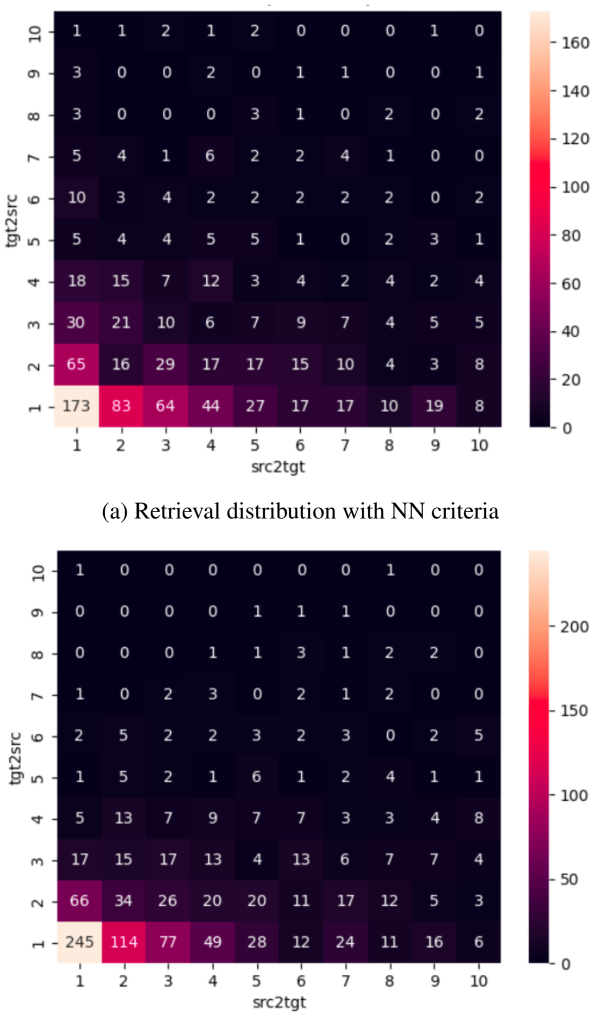
Dataset	Scores		
	@1	@5	@10
Smith et al. (2016): On their original eval dataset[*]	22	40	45
Smith et al. (2016)+NN: On our eval dataset[†]	25	**44**	50
Smith et al. (2016)+CSLS: On our eval dataset[†]	**26**	43	49
our work best results	20	42	**51**

Table 5: Si→En Embedding Alignment Results with previous alignment work
[*] From Smith et al. (2016) official repository [†] Aligned using alignment matrix given in Smith et al. (2016) official repository and evaluated using our evaluation set. The scores can be overestimated since we do not know the exact alignment dataset used by the authors. If there is an intersection between the alignment dataset and our evaluation dataset, the scores may not represent the exact alignment accuracy.

Dataset	Unique Src within 200k	Retrieval					
		NN			CSLS		
		@1	@5	@10	@1	@5	@10
En-Si-para-wiki-5k	5000	11.4	26.4	33.2	14.8	31.5	39.8
En-Si-para-wiki-full	27846	**17.0**	**36.1**	**45.1**	**20.2**	**42.4**	**50.9**
En-Si-para-cc-5k	5000	16.4	35.7	43.6	20.4	39.9	49.1
En-Si-para-cc-full	27856	**17.4**	**37.9**	**45.5**	**20.9**	**42.4**	**50.8**

Table 6: En→Si Procrustes Embedding Alignment Results with different dataset sizes

created our alignment dataset was using the English column of the En-Es MUSE (Lample et al., 2018) alignment datasets and, therefore even if the frequent English words are included, no frequent word selection criterion was imposed on the Sinhala word selection. We assumed that by selecting



(a) Retrieval distribution with NN criteria

(b) Retrieval distribution with CSLS criteria

Figure 1: Top-k Retrieval distribution for RCSL alignment. (The numbers indicate how many pairs in the test set are retrieved in En→Si and Si→En directions with corresponding top-k values)

the most frequent English words would indirectly lead to the most common Sinhala words. Also assumed that MUSE datasets have been created considering the most frequent words in the vocabularies (Lample et al., 2018).

5.3 Alignment Techniques

Where we do not find a proper alignment dataset, we can go for semi-supervised or unsupervised alignment techniques. The unsupervised techniques by Lample et al. (2018) and Grave et al. (2019) have shown competitive results with the supervised techniques. Therefore our next immediate focus will be on semi-supervised and unsupervised alignment techniques.

6 Conclusion

The alignment dataset we used (En-Si-para-cc) has been constructed using the most frequent words in both languages as discussed in Section 5.2. We observed that when we do the alignment using infrequent words (i.e. alignment dictionary created without specifically considering frequent terms) the precision is worse. That is because the most frequent words' embeddings can be assumed well positioned in the embedding spaces rather than infrequent words. That observation has been reported by Smith et al. (2016) as well.

The obtained results show that Si→En alignment is better than EN→Si alignment. We can explain that observation as follows. FatText English embedding space (wiki-256k, cc-2M) is considerably larger than the Sinhala embedding space (wiki-79k, cc-808k). Therefore aligning a larger embedding space onto a smaller space is lossy than the other way around given the probability of a given candidate word from the source not existing in the target is high. Further, given that Sinhala is a highly inflected language compared to English (de Silva, 2019), multiple morphological forms which exist in Sinhala, would invariably map to the parallel of the root word in English. Thus extenuating the viable pool of the Sinhala vocabulary to be matched to their English counterparts. We can assume that these are the reasons contributing to the drop in the resultant improvement of the @5 and @10 precision in En→Si direction during the refinement procedure.

When it comes to the retrieval criterion, the CSLS gives better results than NN in most cases. Then, as far as the training objective is considered, RCSLS with CSLS as the retriever criterion has shown the best precision in most cases. This is because the core idea of RCSL alignment is to make both the training and retrieval consistent rather than using two different criteria (Joulin et al., 2018a). According to Aboagye et al. (2022), the *RCSL* approach by Joulin et al. (2018a) has the highest average alignment quality/accuracy among available cross-lingual embedding alignment techniques and, from our experiments for En-Si alignment, we could verify that fact. We have used alignment datasets with 5k unique source words for the experiments since most of the other work has been carried out with that configuration (Joulin et al., 2018a) but, from Table 6 results we see that we can achieve better results by having a larger dataset.

References

Prince O Aboagye, Yan Zheng, Michael Yeh, Junpeng Wang, Zhongfang Zhuang, Huiyuan Chen, Liang Wang, Wei Zhang, and Jeff Phillips. 2022. Quantized Wasserstein Procrustes alignment of word embedding spaces. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 200–214, Orlando, USA. Association for Machine Translation in the Americas.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *NIPS*, 32.

Nisansa de Silva. 2019. Survey on publicly available sinhala natural language processing tools and research. *arXiv preprint arXiv:1906.02358*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 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)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Riyafa Abdul Hameed, Nadeeshani Pathirennehelage, Anusha Ihalapathirana, Maryam Ziyad Mohamed, Surangika Ranathunga, Sanath Jayasena, Gihan Dias, and Sandareka Fernando. 2016. Automatic creation of a sentence aligned sinhala-tamil parallel corpus. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 124–132.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018a. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018b. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *EMNLP*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

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

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Dimuthu Lakmal, Surangika Ranathunga, Saman Peramuna, and Indu Herath. 2020. Word embedding evaluation for Sinhala. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1874–1881, Marseille, France. European Language Resources Association.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *ICLR*.

Anushika Liyanage, Surangika Ranathunga, and Sanath Jayasena. 2021. Bilingual lexical induction for sinhala-english using cross lingual embedding spaces. In *2021 Moratuwa Engineering Research Conference (MERCon)*, pages 579–584.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2016. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *International Conference on Learning Representations*.

Jörg Tiedemann. 2016. Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3518–3522, Portorož, Slovenia. European Language Resources Association (ELRA).

Charangan Vasantharajan, Laksika Tharmalingam, and Uthayasanker Thayasivam. 2022. Adapting the tesseract open-source ocr engine for tamil and sinhala legacy fonts and creating a parallel corpus for tamil-sinhala-english. In *2022 International Conference on Asian Language Processing (IALP)*, pages 143–149. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NIPS*, 30.

Asanka Wasala and Ruvan Weerasinghe. 2008. Ensitip: a tool to unlock the english web. In *11th international conference on humans and computers, Nagaoka University of Technology, Japan*, pages 20–23.

Kasun Wickramasinghe and Nisansa De Silva. 2023. Sinhala-english parallel word dictionary dataset. In *2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS)*, pages 61–66.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Method	Joulin et al. (2018a)										Ours	
	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-zh	zh-en	en-si	si-en
Adv.+refine+NN	79.1	78.1	78.1	78.2	71.3	69.6	37.3	45.3	30.9	21.9	-	-
Adv.+refine+CSLS	81.7	83.3	82.3	82.1	74.0	72.2	44.0	59.1	32.5	31.4	-	-
Procrustes+NN	77.4	77.3	74.9	76.1	68.4	67.7	47.0	58.2	40.6	30.2	16.4	21.3
Procrustes+CSLS	81.4	82.9	81.1	82.4	73.5	72.4	51.7	63.7	42.7	36.7	20.4	18.0
RCSLS+NN	81.1	84.9	80.5	80.5	75.0	72.3	55.3	67.1	43.6	40.1	21.5	**23.3**
RCSLS+CSLS	**84.1**	**86.3**	**83.3**	**84.1**	**79.1**	**76.3**	**57.9**	**67.2**	**45.9**	**46.4**	**22.6**	19.4

Table 7: Extended Comparison nearest neighbour (NN) and CSLS retrieval Criteria. Here only the top-1 precision scores have been included

A Impact of Retrieval Criterion

Table 7 shows a comparison of how Si-En aligned embeddings behave with different retrieval criteria with other language pairs. In all the other language pair results given in Joulin et al. (2018b), the RCSLS criterion outperforms the NN criterion in both directions but, in our case, Si→En direction, NN has shown the best results while En→Si shows the best results with CSLS. This effect can be clearly seen in Table 3 as well. Joulin et al. (2018b) says, *RCSLS transfers some local information encoded in the CSLS criterion to the dot product.* to establish a suggestion as to why RCSLS outperforms NN in their results but, it seems RCSLS need not be the best retriever criterion for all the cases and, could depend on the language pair and the alignment direction.