Knowledge and Information Systems
https://doi.org/10.1007/s10115-025-02520-4

RESEARCH

Linguistic entity masking to improve cross-lingual representation of multilingual language models for low-resource languages

Aloka Fernando[1] · **Surangika Ranathunga**[2]

Received: 12 May 2024 / Revised: 21 April 2025 / Accepted: 8 June 2025

Abstract
Multilingual Pre-trained Language models (multiPLMs), trained on the Masked Language Modelling (MLM) objective are commonly being used for cross-lingual tasks such as bitext mining. However, the performance of these models is still suboptimal for low-resource languages (LRLs). To improve the language representation of a given multiPLM, it is possible to further pre-train it. This is known as continual pre-training. Previous research has shown that continual pre-training with MLM and subsequently with Translation Language Modelling (TLM) improves the cross-lingual representation of multiPLMs. However, during masking, both MLM and TLM give equal weight to all tokens in the input sequence, irrespective of the linguistic properties of the tokens. In this paper, we introduce a novel masking strategy, *Linguistic Entity Masking* (LEM) to be used in the continual pre-training step to further improve the cross-lingual representations of existing multiPLMs. In contrast to MLM and TLM, LEM limits masking to the linguistic entity types nouns, verbs and named entities, which hold a higher prominence in a sentence. Secondly, we limit masking to a single token within the linguistic entity span thus keeping more context, whereas, in MLM and TLM, tokens are masked randomly. We evaluate the effectiveness of LEM using three downstream tasks, namely bitext mining, parallel data curation and code-mixed sentiment analysis using three low-resource language pairs English-Sinhala, English-Tamil, and Sinhala-Tamil. Experiment results show that continually pre-training a multiPLM with LEM outperforms a multiPLM continually pre-trained with MLM+TLM for all three tasks.

Keywords Masked language modelling · Translation language modelling · Multilingual pre-trained language model · Bitext mining · Sentiment analysis · XLM-R · Sinhala · Tamil

✉ Aloka Fernando
alokaf@cse.mrt.ac.lk

✉ Surangika Ranathunga
s.ranathunga@massey.ac.nz

[1] Department of Computer Science & Engineering, University of Moratuwa, Moratuwa 10400, Sri Lanka

[2] School of Mathematical and Computational Sciences, Massey University, Palmerston North 4443, New Zealand

Published online: 19 July 2025

1 Introduction

Encoder-based Multilingual Pre-trained Language Models (multiPLMs) such as mBERT [1] and XLM-R [2] produce State-of-the-art results for many Natural Language Processing (NLP) tasks in the context of low-resource languages (LRLs) [3, 4]. One success factor of these multiPLMs is the training objective utilized during the pre-training stage. mBERT and XLM-R were pre-trained using the Masked Language Modelling (MLM) objective. However, the performance of these models for cross-lingual tasks such as bitext mining had been suboptimal [5], due to the lack of an explicit cross-lingual pre-training objective [6]. Translation Language Modelling (TLM) objective [7] was introduced to improve the cross-lingual capability of the existing multiPLMs. In contrast to MLM that uses only monolingual data, TLM uses parallel data across multiple languages in a *continual pre-training* step to further improve the cross-lingual representations. Conneau and Lample [7] had proven that TLM improved the performance of cross-lingual classification and unsupervised Neural Machine Translation (NMT).

From a linguistic perspective, different words in a sentence have different linguistic properties. Previous work has demonstrated that Pre-trained Language Models (PLMs) capture the notion of syntactic structures and grammatical properties in the language [8–10]. Named Entities (NEs), Verbs and Nouns significantly contribute to defining the syntactic structure and the semantics of the sentence. Further, these elements play a crucial role in establishing syntactic relationships such as subject-verb agreement, which is generally stronger than those between other words in the sentence. To highlight the prominence of NEs, Verbs and Nouns in a sentence, we visualize the self-attention weight matrix in terms of a heatmap for an English sentence *Jack walks towards the road*, and its Sinhala translation in Fig. 1. In the English sentence, the words "Jack" (NE) and "walk" (Verb) get the highest attention from other words. Similarly, the words which gets the highest attention in Sinhala is a Named Entity and a Noun.

Based on this hypothesis, we introduce a linguistically motivated masking strategy named *Linguistic Entity Masking (LEM)*. In LEM, we limit to masking a single token from the linguistic entity span. As linguistic entities, we consider NEs, nouns and verbs in the sentence. Masking a single token contrasts existing span masking techniques [11–13], which have masked consecutive token spans of the selected n-gram words. We apply LEM on both

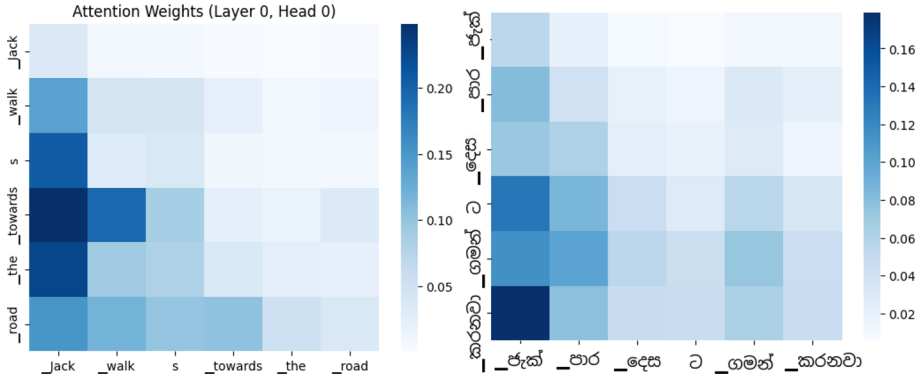

Fig. 1 Self-attention weights among the words for an English and its corresponding Sinhala sentence. The darker the colour is, the stronger the relationship (ie. self-attention weight) between the two words

monolingual (LEM_{mono}, akin to MLM) and parallel sentences (LEM_{para}, akin to TLM), and conduct continual pre-training on the multiPLM. We use XLM-R as the multiPLM in our experiments. We continually pre-train XLM-R with $LEM_{\text{mono}} + LEM_{\text{para}}$ objectives. The resulting model is referred to as XLM-R_{LEM} hereafter. We continually pre-train this same multiPLM with MLM+TLM as well, which serves as our baseline (we term the resulting model XLM-R$_{MLM+TLM}$). We evaluate these models on bitext mining, parallel data curation, as well as code-mixed sentiment classification downstream tasks on three LRL pairs English-Sinhala, English-Tamil, and Sinhala-Tamil. Our results showcase that XLM-R_{LEM} outperforms XLM-R$_{MLM+TLM}$ for these tasks.

We carry out extensive experiments (1) to determine the type of monolingual data that is most effective in the first LEM_{mono} continual pre-training step. ie. *dependent monolingual data* (source and target sides of parallel data taken separately as monolingual data) vs *independent monolingual data* (monolingual data without any explicit translation content between the two languages), (2) to empirically evaluate the existing masking strategies (Sub-word masking [1], whole-word masking [1] and span Masking [12]), (3) to identify the most contributing linguistic entity or combination of linguistic entities for masking, (4) to determine the optimal number of tokens for masking within a linguistic entity and (5) to determine the impact of using noisy parallel sentences in the continual pre-training stage. Our contributions are as follows:

- We introduce a novel masking strategy, Linguistic Entity Masking (LEM), to improve the cross-lingual representations of existing multiPLMs. We show that XLM-R_{LEM} outperforms XLM-R$_{MLM+TLM}$ for the three downstream tasks, considering English and two LRLs Sinhala and Tamil.
- We conduct an empirical study of the existing masking strategies to evaluate the effectiveness of them on the bitext mining task for LRLs.
- We conduct ablation experiments to find the most contributing linguistic entity (among NEs, nouns and verbs) for masking and the number of tokens to mask within the linguistic entity span.
- We show that using dependent monolingual corpora than the independent monolingual corpora during the continual pre-training step is favourable for improving the cross-lingual representations.

The rest of the paper is organized as follows. In Sect. 2 we discuss the related work in the context of pre-training objectives and masking strategies. In Sect. 3 we describe our methodology, assumptions along with the theoretical justification for the LEM strategy. We describe the experiments in Sect. 4 and detail out the experimental setup in Sect. 5. The results are discussed comprehensively in Sect. 6, while additional ablation studies are detailed in Sect. 7. The limitations of our approach and potential future directions are highlighted in the Discussion Sect. 8, and the paper conclusion is in Sect. 9.

2 Related work

2.1 MLM and TLM objectives

Encoder-based multiPLMs such as mBERT and XLM-R were trained on monolingual data using the Masked Language Modelling (MLM) objective. These models have significantly enhanced the performance of various downstream tasks [14–16].

In BERT (and its multiPLM variant, mBERT), which was trained with MLM,[1] 15% of the input tokens were randomly selected for corrupting, following a uniform distribution. Out of these, 80% of the time the tokens were replaced with a [MASK] token, 10% of the time the tokens were replaced with a random token and 10% of the time they were left unchanged. Here the MLM objective predicts the corrupted (both masked and replaced) tokens. Successor models such as XLM-R [2] adopt the same 15% masking percentage and 80%-10%-10% corruption rule during pre-training. The contextualized representations produced by these pre-trained models are then used to obtain sentence embeddings for downstream NLP tasks. However, these models have been reported to be suboptimal for cross-lingual tasks such as bitext mining, due to the lack of an explicit objective for improving cross-lingual representations [6].

To address this limitation, Conneau and Lample [7] introduced Translation Language Modelling (TLM), which extended the MLM objective using parallel data. TLM accepts a concatenated pair of parallel sentences as input, and tokens were masked from both sentences. The rationale was to utilize the context of its translation counterpart to accurately predict the masked token, there by strengthening the cross-lingual capability. TLM was applied in a continual pre-training step, on top of the MLM pre-trained model. In this setting, the MLM step was still required to learn the linguistic information inherent to the languages, while the TLM step strengthened the cross-lingual signal across the language pairs.

2.2 Different masking strategies

In BERT's MLM strategy, the sub-words were masked. Subsequent work experimented by varying the type of tokens for masking, as summarized in Table 1. The follow-up work masked consecutive sub-words in text spans [12] and correlated text spans with Point-wise Mutual Information (PMI) masking [13]. Zhuang [17] proposed a heuristic-masking strategy where they considered the unmasked token prediction in addition to the masked token prediction during language model pre-training. Golchin et al. [18] utilized an in-domain keyword masking strategy for domain adaptation of the PLM. Most of these techniques have primarily relied on monolingual data and have been evaluated predominantly on high-resource languages.

Closest to our work is Entity/Phrase masking [11]. This contrasts with our work in three ways. Entity/Phrase masking, selected NEs, noun *phrases* and verb *phrases* for masking. As per analysis in Fig 1, we considered only verb and noun *words*, in addition to NEs for masking. Secondly, while Entity/Phrase masking masked all consecutive tokens within NEs or Noun/Verb *phrases* identified by a chunking tool, LEM takes a more targeted approach by limiting masking to a single token within the selected linguistic entity span. Finally, they followed a multi-staged pre-training approach. The pre-training stages were, sub-word masking similar to BERT, followed by Phrase masking and Named Entity masking. In comparison, our approach's two continual pre-training stages apply the same LEM strategy with monolingual data and parallel data. They have evaluated the strategy only on High Resource Languages English and Chinese. Further, this strategy has not been extended with parallel data for cross-lingual improvement.

Wettig et al. [19] conducted an empirical study on what tokens should be masked and in what percentages. However, their study was limited to the English language and focused only on downstream tasks such as classification and question-answering. To date, no empirical study has explored these alternative masking strategies specifically for sentence retrieval tasks, particularly in the context of LRL pairs.

[1] BERT was also trained using the next sentence prediction task.

Table 1 Existing masking strategies.

Masking Strategy	Pre-training	Masked token Type
Sub-word Masking	Pre-training	Sub-words
Whole-Word Masking	Pre-training	All sub-words in the word
Entity/Phrase Masking [11]	Multi-stage Pre-training	All sub-words in the Named Entity/Noun Phrase
Span Masking (spanBERT) [12]	Pre-training	All sub-words in the word n-gram span
Point-wise Mutual Information (PMI) Masking [13]	Pre-training	all sub-words in the correlated word-spans

The *Masked Token Type* indicates the type of words considered for masking

Fig. 2 A comparison of the existing masking strategies considering an example from the English-Sinhala language pair. Sub-word masking, Whole Word masking, span masking, and LEM_{mono} consider only monolingual sentences during masking. TLM and LEM_{para} consider concatenated parallel sentences to apply the masking. In LEM_{mono} and LEM_{para}, only a single token from the linguistic entity is masked

3 Methodology

In this section, we discuss in detail the LEM strategy. A comparison between our masking strategy and existing masking strategies is shown in Fig. 2. Instead of pre-training a multiPLM from scratch-a computationally expensive process-we leverage LEM in a continual pre-training step. This is a widely adopted approach to improve multiPLMs with respective to representation improvements [7, 20].

Figure 3 illustrates our two-stage continual pre-training process. Similar to the MLM and TLM training sequence used in XLM [7], on top of the multiPLM, we apply the continual pre-training with monolingual and parallel data, respectively.

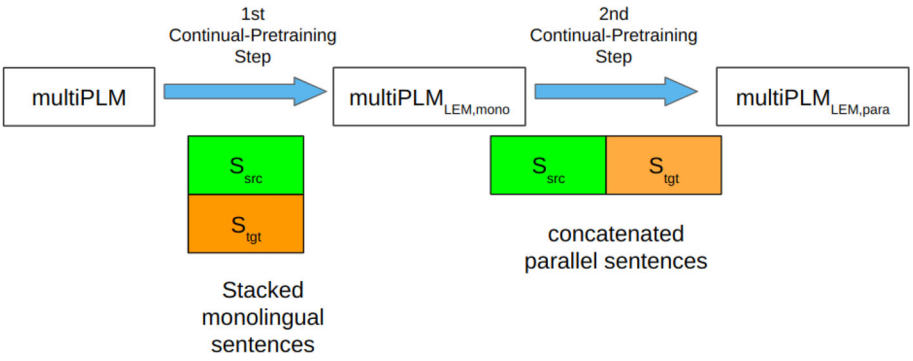

Fig. 3 The LEM continual pre-training process. As *multiPLM*, we select an existing multilingual pre-trained language model. The first step ie. LEM$_{mono}$ is to continually pre-train with *stacked monolingual sentences*, meaning the monolingual data from the source side is passed first, followed up by the target language monolingual data. In the second continual pre-training step ie. LEM$_{para}$, the LEM strategy is applied on the *concatenated parallel data*

3.1 Theoretical framework for linguistic entity masking (LEM)

The theoretical framework of LEM in a monolingual setting (LEM_{mono}) can be described as follows:

Let the monolingual sequence X be defined as $X = x_1\ x_2\ x_3\ ...x_i...\ x_n$ where x_i is a word and n is the number of words in the sequence. After tokenization, sequence X can be represented as $\bar{X}$ as in Eq. 1.

$$\bar{X} = \bar{x}_1\ \bar{x}_2\ \bar{x}_3\ \bar{x}_4 ..... \bar{x}_j .... \bar{x}_m \tag{1}$$

Here, $\bar{x}_j$ is a token (sub-word) and m is the total number of sub-words returned by the tokenizer. From this sequence, the linguistic entities NEs, verbs and nouns are identified, and $\bar{X}$ can now be represented as a collection of linguistic entities as shown in Eq. 2. From these linguistic entities, a single token is sampled over a uniform distribution, up to a total of 15% for masking. If 15% cannot be obtained from linguistic entities, the remainder would be sampled from the remaining tokens. We use the same corruption rule, 80%-10%-10% as BERT.

$$\bar{X} = \{\{\bar{x}_1\ \bar{x}_2\}, ...\ \{\bar{x}_4\bar{x}_5\bar{x}_6\}, ...\{\bar{x}_m\}\} \tag{2}$$

During training, the cross-entropy loss ($\mathcal{L}_{LEM_{mono}}$) for masked token prediction, as in Eq. 3 is minimized. In the equation, N is the total number of tokens for prediction and y_j is the expected true label.

$$\mathcal{L}_{LEM_{mono}} = -\frac{1}{N}\sum_{j=1}^{N} y_j \log(P(x_j)) \tag{3}$$

Finally, we extend the TLM objective with LEM into the parallel data setting (LEM$_{para}$). Here we concatenate the source sentence ($X = x_1\ x_2\ x_3\ .....\ x_m$) and target sentence ($Y = y_1\ y_2\ y_3\ .....\ y_n$) from the parallel sentence-pair as a single input example and obtain the tokenized output as represented by $\bar{Z}$ in Eq. 4. $\bar{X} = \bar{x}_1\ \bar{x}_2\ \bar{x}_3 ..... \bar{x}_k$ and $\bar{Y} = \bar{y}_1\ \bar{y}_2\ \bar{y}_3 ..... \bar{y}_l$ are the tokenized source and target sentences, respectively. k and l are the number of tokens (sub-words) in the source and target sentences (respectively) after tokenization.

$$\bar{Z} = \bar{x}_1\ \bar{x}_2\ \bar{x}_3 ..... \bar{x}_k\ \bar{y}_1\ \bar{y}_2\ \bar{y}_3 ..... \bar{y}_l \tag{4}$$

Table 2 English (En), Sinhala (Si) and Tamil (Ta) examples of the returned sub-words after the tokenization step. English nouns get inflected based on the number only. But Sinhala and Tamil nouns get inflected based on case and gender as well

Type	singular/subject	plural/subject	plural/object	singular/feminine
Original word (Si)	ගුරුවරයා			
Si word (inflected)	ගුරුවරයා	ගුරුවරු	ගුරුවරුන්	ගුරුවරිය
Tokenized output	ගුරු", "#වරයා"	ගුරු", "#වරු"	ගුරු", "#වරුන්",	"ගුරු", "#වරිය"
Original word (En)	Teacher			
En word	teacher	teachers	teachers	the teacher
Tokenized output	"teacher"	"teacher", "#s"	"teacher", "#s"	"the", "teacher"
Original word (Ta)	ஆசிரியர்			
Ta word (inflected)	ஆசிரியர்	ஆசிரியர்கள்	ஆசிரியர்கள்	ஆசிரியர்
Tokenized output	"ஆசிரியர்"	"ஆசிரியர்", "#கள்"	"ஆசிரியர்", "#கள்"	"ஆசிரியர்"

Similar to LEM_{mono}, in this step, a single token from each linguistic entity (NE, verb or noun) from both languages are selected for corruption according to the 80%-10%-10% rule. If 15% of linguistic units were not found in the sequence, the balance is sampled from the remaining tokens. During training, the corrupted token prediction cross-entropy loss, ($\mathcal{L}_{LEM_{para}}$) (Eq. 5) is minimized. S and T correspond to the total number of tokens masked from the source and target side sentences, respectively. z_s and z_t are the true tokens to be predicted.

$$\mathcal{L}_{LEM_{para}} = -\frac{1}{S}\sum_{s=1}^{S} z_s \log(P(x_s) - \frac{1}{T}\sum_{t=1}^{T} z_t \log(P(j_t)) \tag{5}$$

Languages such as Sinhala and Tamil exhibit morphological richness, requiring words to be inflected based on attributes, such as number, gender, and case category. Table 2 shows examples for such word inflections. Additionally, the presence of out-of-vocabulary words in LRLs often leads to an increased number of sub-words after tokenization. Therefore, approaches like whole-word masking, span masking, or entity/phrase masking tend to mask longer spans. This reduction in context weakens the ability to accurately predict the masked tokens, ultimately hindering representation learning. In contrast, LEM mitigates this issue by masking a single token from a linguistic entity, which we empirically prove in Sect. 7.1.

4 Experiments

4.1 Impact of the type of monolingual data in LEM_{mono}

We experiment with independent and dependent monolingual data to observe its impact on the continual pre-training step LEM_{mono}. We sample 60,000 sentences from the MADLAD-400 and all the available 60,000 sentences from SiTa-Trilingual dataset for each language. Next, we continually pre-train XLM-R with those datasets separately with LEM_{mono} and evaluate on the bitext evaluation dataset.

To assess whether increasing the independent training data would yield in any improvements, we repeat the experiment with a sample size of 100,000 from MADLAD-400. Finally,

as an extreme case, we conduct a third experiment using 500,000 sentences for the Sinhala-Tamil language pair. Here, 60,000, 100,000, and 500,000 correspond to the training data size per language, with the full training set size being double the amount specified. We conduct this evaluation for all three language pairs.

4.2 Evaluation of different masking strategies

We empirically evaluate various masking strategies and assess their performance on the bitext mining task. The masking strategies explored in this study are as follows:

Sub-word Masking - Following the BERT MLM, with each sentence, 15% of tokens are selected randomly and corrupted according to 80%-10%-10% rule.

Whole Word Masking - All the sub-words corresponding to the randomly sampled words are masked. A total of 15% tokens are sampled and corrupted according to 80%-10%-10% rule.

Span Masking - Consecutive word spans are sampled over a geometrical distribution and 15% of tokens are masked. The masking is limited to whole-word tokens as defined in the original work.

4.3 Evaluation of LEM strategy and ablation study

This section describes the ablation experiments we conduct to determine the most contributing linguistic entity or their combination in the LEM strategy. We use the baselines as described in Sect. 5.3.

We identify the NEs in English, Sinhala, and Tamil sentences, using an in-house fine-tuned multilingual NER model [21]. To identify nouns and verbs in the sequences, we employ Flair POS tagger [22] for English, the Sinhala TnT POS Tagger [23, 24] for Sinhala, and ThamizhiUDp [25] for Tamil. Flair reported an F1 score of 98.19% and is the best model for English POS Tagging. The Sinhala POS Tagger had been trained using SVM and has an overall accuracy of 84.68% with a 59.86% accuracy for tagging unknown words. The Tamil POS Tagger is a neural-based model, with a F1 score of 93.27%. For Sinhala and Tamil, these are the models that returned optimal results.

The initial ablation experiment masks a single linguistic entity type, such as only NEs, only verbs, or only nouns. Subsequently, combinations of these linguistic entity types are examined.

In our experiments, 100%NE+15%MLM means, priority is given for sampling from NEs. If it does not produce enough tokens for masking, then the balance is sampled from the remaining tokens. When combining several linguistic entities, e.g. 100%NE+100%VB+15%MLM means, priority is given to sample the tokens for masking from both NEs and verbs.

4.4 Evaluation tasks

We evaluate the success of our LEM masking strategy on three downstream tasks - bitext mining, parallel data curation and code-mixed sentiment classification.

4.4.1 Bitext mining

Bitext mining is a sentence retrieval task that retrieves a target language translation for a given source sentence or vice versa from a document-aligned dataset. The performance of bitext mining relies heavily on the quality of cross-lingual embeddings. Recent bitext mining techniques are embedding-based, where they identify the translation pairs based on the semantic distance between the source sentences and candidate target sentences. Improvements in bitext mining techniques can be categorized into two: (1) refining the semantic similarity distance calculation function [26, 27] between sentence embeddings and (2) enhancing cross-lingual sentence representations [20, 28, 29]. Our LEM strategy aims to improve the latter.

We obtain the sentence embeddings from XLM-R$_{LEM}$ as well as from the baselines, and use margin-based cosine similarity [26] function to identify parallel sentence pairs. We choose margin-based cosine similarity over conventional cosine similarity for this task due to its lower rate of false positives. Then we rank the parallel sentences according to their similarity scores. Bitext mining is performed using the three criteria: Forward (FW), Backward (BW), and Intersection (IN) [26]. FW retrieves the target sentence for each source sentence, BW retrieves the source sentence for each target sentence, and IN considers the intersection of the parallel sentences retrieved using FW and BW criteria. The Recall evaluation metric is used to report the bitext mining results.

4.4.2 Parallel data curation

Although large-scale bitext mining [30, 31] alleviates the parallel data scarcity problem in NMT, they are mostly noisy [32, 33]. Therefore, a parallel data curation step is crucial to filter out noisy parallel sentences from the corpus. This is done by obtaining sentence representations from a multiPLM and by calculating the semantic similarity using cosine distance [20, 31] between each parallel sentence pair. Then the parallel sentences are sorted in descending order and the top-most ranked parallel sentences are used to train the NMT system.

To further evaluate the effectiveness of these models with improved cross-lingual representations, we conduct a parallel corpus curation task and perform an extrinsic evaluation by training NMT systems with the top-ranked sentences.

First, we rank the parallel sentences for translation quality using the baseline (Sect. 5.3) and the XLM-R$_{LEM}$ models. Using the top 50,000 sentences from the ranked parallel sentence pairs, we train NMT systems for each language pair. NMT scores are reported on the Flores+ devtest, using sacreBLEU [34], ChrF [35], ChrF++ [36] and spBLEU [37] metrics. We base the discussion of the results using the ChrF metric while we include the full results Table 13 in Appendix 1.

4.4.3 Code-mixed sentiment analysis

MultiPLMs fine-tuned on task-specific monolingual datasets has achieved the state-of-the-art performance in sentiment analysis tasks [38]. However, sentiment analysis on code-mixed data remains a challenging task. Code-mixing [39] occurs when linguistic units-such as phrases, words, or morphemes-from one language are embedded into the utterance or sentence of another language. In this setting, the performance of sentiment analysis on code-mixed data largely depends on the quality of the cross-lingual embeddings learned by the MultiPLM.

We conduct the baseline experiments (Sect. 5.3) and compare them with the results we obtain with the XLM-R$_{LEM}$ model. Here we use each encoder model by fine-tuning them on

an English-Sinhala code-mixed task. We follow a two-step fine-tuning, where the first fine-tuning is done using the English Amazon product review sentiment analysis dataset. Here we report the zero-shot scores on the code-mixed EnSi evaluation set. Finally, the intermediate model is further fine-tuned on the English-Sinhala code-mixed dataset and the results are reported. We use Precision, Recall and F1 to report the evaluation scores for the sentiment analysis task.

5 Experiment setup

5.1 Data selection

We consider English, Sinhala and Tamil for our experiments. Sinhala and Tamil are the official languages of Sri Lanka and English is used as a link language. They belong to three distinct language families; Indo-European, Indo-Aryan and Dravidian language families, respectively. Further, Sinhala and Tamil are low-resource and medium-resource languages, respectively [40, 41]. They are also morphologically rich languages. Sinhala, in particular, is only used in the island nation of Sri Lanka and has seen slow progress in language technologies [41, 42].

Monolingual and Parallel Data: As elaborated in Sect. 4.1, we carry out an ablation study to determine the type of monolingual data that is most suitable for the first continual pre-training step. We obtain the independent monolingual data from MADLAD-400 [43]. It is a collection of document-level data of 3 Trillion tokens from Common Crawl[2] for 419 languages. As the dependent-monolingual data, we obtain the monolingual sides from the SiTa-Trilingual parallel dataset [44]. It is a human-curated gold standard three-way parallel dataset between Sinhala-Tamil-English languages with 60,000 training data.

We preprocess the MADLAD-400 data to extract clean sentences for each language as follows: First, the document-level data is segmented into sentences using the nltk[3] sentence tokenizer. We then filter these sentences using the LID (**L**anguage **ID**entification) model.[4] Subsequently, we remove noisy data, including HTML tags, URLs, and sentences with less than 60% textual content. Finally, religious texts were excluded through a keyword filter.[5] However, for SiTa-Trilingual data, no preprocessing was applied as the data was of high quality.

NLLB/CCAligned Datasets: For the parallel data curation task, we obtain parallel data from NLLB [31] and CCAligned [45] corpora. Both these corpora provide parallel data for the three language pairs: English-Sinhala, English-Tamil, and Sinhala-Tamil. NLLB and CCAligned are known to be noisy parallel data for the considered language pairs [33].

ParaCrawl Dataset: To analyse the performance of LEM strategy with noisy data, we select the English-Sinhala ParaCrawl [46] dataset. It is a web-mined parallel corpus with 217,412 parallel sentences.

[2] https://commoncrawl.org/.

[3] https://www.nltk.org/index.html.

[4] https://github.com/gordicaleksa/Open-NLLB.

[5] Keywords being bible book names along with common words from the bible.

Code-mixed Sentiment Classification pairsDataset: For the code-mixed sentiment classification task, we use the English-Sinhala code-mixed dataset [47] of 13,521 sentences. We have experimented with the English-Sinhala language pair.

Amazon Product Review Dataset: To report the zero-shot scores for the English-Sinhala sentiment classification task, we used the Amazon product review sentiment analysis dataset .[6] The full dataset has training data of 3.6M and test data 400,000, respectively. During fine-tuning, we sample only 100,000 as training data, in order to avoid catastrophic forgetting [48] of the cross-lingual representations in XLM-R_{LEM}

Trilingual Bitext Mining Evaluation Set: For the bitext mining task, we use an existing human-created dataset [27]. It consists of trilingual data obtained from four Sri Lankan news sources Army,[7] Hiru,[8] ITN[9] and Newsfirst.[10] For each news source, there are human-aligned 300 sentence-pairs.

Flores+ Evaluation Set: For the NMT experiments, we use dev and devtest splits from Flores+[11] as the validation and evaluation sets, respectively.

5.2 multiPLM selection

We select XLM-R as the base multiPLM for our experiments. Other popular multiPLMs, XLM and mBERT do not cover Sinhala language. XLM-R has already demonstrated promising performance in downstream tasks for the low-resource languages considered in this study [4, 21, 47, 49]. It is a 278 M parameter model, pre-trained for 100 languages. However, the amount of Sinhala and Tamil data used during XLM-R pre-training is much lower than that for English (English 55B, Sinhala 243 M, Tamil 595 M).

5.3 Baselines

In our evaluation of downstream tasks, we set up two baseline experiments.

- **XLM-R** - Obtain embeddings from the out-of-the-box XLM-R pre-trained model.
- **XLM-R**$_{MLM+TLM}$ [7] - We continually pre-train the XLM-R with MLM+TLM objectives and use the representation improved encoder to obtain embeddings.

5.4 Implementation and Hyper-parameters

5.4.1 Linguistic entity masking (LEM)

We customize the MLM training implementation released with the sentence-transformers[12] library (built on Huggingface transformers[13]), to support XLM-R tokenization and implement

[6] https://www.kaggle.com/datafiniti/consumer-reviews-of-amazon-products.

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

[8] https://www.hirunews.lk/.

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

[10] https://english.newsfirst.lk/.

[11] https://github.com/openlanguagedata/flores.

[12] https://www.sbert.net/.

[13] https://huggingface.co/docs/transformers/index.

Table 3 Hyper-parameters used during continual pertaining with LEM strategy

Hyperparameter	Argument value
No of Layers	12
Hidden Size	768
Attention Heads	12
hidden_dropout_prob	0.1
Learning Rate	5e-3
Training batch-size	32
Sequence Length	120
Adam ϵ	1 e-08
Adam β_1	0.9
Adam β_2	0.99

the LEM strategy. Each continual pre-training experiment is executed for 60 epochs with early stopping. Then the checkpoint with the least validation loss is selected as the best-performing model. The experiments are conducted on Nvidia Quadro RTX 6000 GPU with 24GB VRAM. The hyper-parameters of XLM-R [14] model and other training parameters used in the continual pre-training experiments are shown in Table 3.

5.4.2 NMT experiments

We obtain the top-ranked sentences for the NMT experiments and train a Sentencepiece[15] tokenizer with a vocabulary size of 25000. Then we use fairseq toolkit [50] to model and train the vanilla transformer-based Sequence-to-Sequence NMT model. The experiments are conducted on a Nvidia Quadro RTX6000 GPU with 24GB VRAM. The hyper-parameters used during training along with the training parameters are shown in Table 4. Each experiment is run for 100 epochs with the early stopping criteria.

5.4.3 Code-mixed sentiment analysis

For the sentiment classification experiments, we fine-tune both the baseline models and the XLM-R$_{LEM}$ model. A linear layer is added as the classification head to facilitate binary sentiment classification. Full fine-tuning is performed, allowing updates to all model parameters, including those in the classification head, during training. The Hyper-parameters used during these experiments are presented in Table 5. We train for 20 epochs with early stopping.

5.4.4 Improving continual pre-training efficiency

Named Entity Recognition (NER) and Part-of-Speech (POS) tagging during training increases training time drastically. We introduce a pre-processing step to mitigate this issue. Specifically, a dictionary is created to store the linguistic entities, such as named entities, verbs, and nouns, for each sentence. We maintain a sub-word-level mapping in the dictionary, allowing for precise token identification while reducing computational overhead during training.

[14] https://huggingface.co/FacebookAI/xlm-roberta-base.

[15] https://github.com/google/sentencepiece.

Table 4 Training parameters for NMT experiments

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

Table 5 Training parameters for the sentiment classification task. experiments

Hyper-parameter	Argument value
Number of Layers	12
Hidden Size	768
Attention Heads	12
Hidden dropout prob	0.1
Learning Rate	2e-5
Weight Decay	0.01
Training Batch Size	128
Sequence Length	80
Adam ϵ	1e-08
Adam β_1	0.9
Adam β_2	0.99

6 Results and discussion

6.1 Impact of the type of monolingual data in *LEM$_{mono}$*

Figure 4 shows the bitext mining results for the LEM_{mono} step using independent and dependent monolingual data. The detailed results are available in Table 12 in Appendix 1.

The highest performance was observed for the dependent monolingual data, a trend consistent across all three language pairs. Surprisingly, when the independent set was increased to 100,000, this increase did not surpass the scores obtained using the dependent monolingual 50,000 training dataset. This was evident in all three language pairs. When the dataset

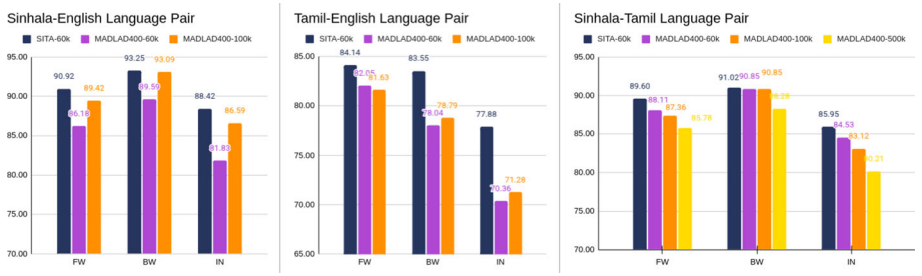

Fig. 4 Bitext mining Recall scores for using independent monolingual data (MADLAD-400) versus dependent monolingual data obtained from the parallel corpus (SiTa-Trilingual parallel Corpus)

size was further increased to 500,000, the results further deteriorated.[16] Therefore, it is evident that utilizing dependent monolingual data is advantageous during the LEM$_{mono}$ step to enhance cross-lingual representations.

6.2 Evaluation of different masking strategies

Table 6 shows the experimental results for the bitext mining task using XLM-R continually pre-trained with different MLM strategies (Section 4.2). Based on the averaged recall scores, the baseline XLM-R consistently delivered the highest performance in bitext mining tasks. An exception was observed for the Sinhala-Tamil BW criterion, where the sub-word masking strategy outperformed the baseline. However, a key observation was that continual pre-training with the existing masking strategies in general deteriorated the already learnt cross-lingual representations in the XLM-R model

As outlined in Sect. 3.1, morphologically rich, low-resource languages tend to generate a higher number of sub-word tokens during the tokenization process due to the presence of infrequent words in the sequences. Consequently, masking strategies like whole-word masking and span masking result in longer masked spans, which reduce the available context for accurate predictions. We hypothesize that this reduced context reduces the prediction accuracy, thus weakening the already learnt representations in the XLM-R model. As a result, the existing masking strategies return degraded results.

6.3 Evaluation of LEM strategy and ablation study

The ablation study results for considering each linguistic entity along with combinations of them, on the LEM strategy are available in Tables 14, 15, 16 in the Appendix 1. In Table 7, we summarize the final gains.

For the Sinhala-Tamil language pair, compared to the scores obtained with XLM-R raw embeddings, XLM-R$_{LEM}$ returned the highest gain of +3.1 Recall points. Compared with XLM-R$_{MLM+TLM}$, this was a gain of +1.4 points. For the English-Tamil language pair, the LEM$_{mono}$ step produced a reduced score compared to the XLM-R baseline. However, after the second continual pre-training step XLM-R$_{LEM}$ scores surpassed the XLM-R baseline by +1.2. Further, the highest gain produced by our method compared to XLM-R$_{MLM+TLM}$ was for the English-Tamil language pair, which was +2.4 points. With the English-Sinhala

[16] Due to resource constraints and the reduced performance for the Sinhala-Tamil direction, the experiment using 500,000 was not conducted for the other two language pairs.

Table 6 Bitext mining recall scores for different masking strategies are shown. The best result for each language pair is highlighted in bold.

Experiment	Army			Hiru			ITN			Newsfirst			Averages		
	FW	BW	IN	FW	BW	IN	FW	BW	IN	FW	BW	IN	FW	BW	IN
English - Sinhala															
XLM-R	**92.33**	93.33	**89.67**	**96.35**	**96.68**	**95.68**	**94.00**	96.00	92.33	**96.67**	**95.33**	**94.33**	**94.84**	**95.34**	**93.00**
Sub-word Masking	88.33	**93.67**	85.33	92.03	93.36	89.70	91.67	**96.67**	**93.67**	91.67	95.33	90.00	90.92	94.76	89.68
Whole-word Masking	87.33	92.67	85.33	95.02	94.01	94.02	93.00	91.67	90.33	93.67	93.67	91.67	92.25	93.00	90.34
Span Masking	89.00	89.67	85.00	95.02	94.02	92.03	90.33	91.67	85.67	93.67	92.67	90.33	92.00	92.01	88.26
English - Tamil															
XLM-R	**86.67**	**88.33**	**82.00**	**83.00**	**78.33**	**72.67**	83.22	**83.56**	**78.86**	**92.33**	**91.33**	89.33	**86.31**	**85.39**	**80.71**
Sub-word Masking	84.00	86.00	77.67	80.33	75.00	68.33	**83.56**	82.21	78.52	90.67	91.00	**89.67**	84.64	83.55	78.55
Whole-word Masking	83.33	87.33	77.67	78.67	73.33	64.33	80.20	80.87	75.84	85.67	91.00	83.67	81.97	83.13	75.38
Span Masking	82.67	83.00	75.33	78.67	76.67	69.33	83.22	82.22	76.85	89.67	90.00	85.67	83.56	82.97	76.79
Sinhala-Tamil															
XLM-R	83.44	81.46	78.15	**90.67**	91.00	**87.33**	**91.33**	90.00	87.00	**93.67**	**95.33**	**92.33**	**89.78**	89.45	**86.20**
Sub-word Masking	**86.75**	88.08	**81.96**	88.00	89.33	84.00	93.33	**92.67**	**89.33**	90.33	94.00	89.00	89.60	**91.02**	86.07
Whole-word Masking	85.76	**89.73**	81.46	88.33	**91.33**	84.67	90.33	90.33	86.67	90.00	91.67	87.67	88.61	90.77	85.11
spanMasking	85.78	85.10	81.79	88.67	91.00	87.00	91.00	91.00	87.33	89.00	90.67	84.33	88.61	89.44	85.11

Table 7 Results for bitext mining task in terms of recall points. For comparison purposes, the FW, BW and IN gains are averaged and reported in the *Overall Average Gain column*

	Average Gains			Overall Average
	FW	BW	IN	Gain
Sinhala-Tamil				
XLM-R$_{LEM}$ vs XLM-R	+2.36	+4.14	+2.90	+3.1
XLM-R$_{LEM}$ vs XLM-R$_{MLM+TLM}$	+1.95	+0.48	+1.83	+1.4
English-Tamil				
XLM-R$_{LEM}$ vs XLM-R	+0.75	+1.59	+1.17	+1.2
XLM-R$_{LEM}$ vs XLM-R$_{MLM+TLM}$	+2.34	+1.84	+2.92	+2.4
English-Sinhala				
XLM-R$_{LEM}$ vs XLM-R	+0.25	+0.50	+0.42	+0.4
XLM-R$_{LEM}$ vs XLM-R$_{MLM+TLM}$	+1.50	+1.50	+2.08	+1.7

language pair, a similar behaviour was observed. Here XLM-R$_{LEM}$ compared to baseline XLM-R has a gain of +0.4 points while the improvement compared to XLM-R$_{MLM+TLM}$ was +1.7 scores.

In all these language pairs, the best scores were produced when the linguistic entity NEs were included in the LEM strategy. Due to the consistent gains across the three language pairs, we can safely conclude that LEM is favourable to improving the cross-lingual representations in existing multiPLMs.

6.4 Parallel data curation

Table 8 shows the NMT results for training the top 50,000 sentence pairs obtained by ranking the parallel sentences with the baseline and XLM-R$_{LEM}$ models for the NLLB and CCAligned corpora. It can be observed consistently that both XLM-R$_{MLM+TLM}$ and XLM-R$_{LEM}$ improved models outperform the baseline XLM-R scores.

The NMT results show that the XLM-R$_{LEM}$ model produce superior results compared to XLM-R$_{MLM+TLM}$, for all three language pairs across the two corpora. We believe the magnitude of the gain is dependent on the characteristics of the parallel corpus and the size of the training data sample. For the English-Tamil language pair, the CCAligned corpus produce a significant gain for XLM-R$_{LEM}$ compared to XLM-R$_{MLM+TLM}$. This justifies the effectiveness of the LEM strategy which was not evident with random masking followed in MLM+TLM. The rest of the gains vary from +0.3 to +0.8 ChrF points. According to metric analysis by Kocmi et al. [51], these gains are equivalent to +0.48 to +1.12 BLEU points with a human accuracy of 54.2% to 66% respectively. This means the improvement in the translation quality in the NMT systems is almost in line with a minimum human accuracy rating of 54.2% to 66%.

The observations are consistent with other metrics, as shown in Table 13 in the Appendix 1 as well. The results further prove that the scoring from the XLM-R$_{LEM}$ model has managed to identify quality sentence pairs more than the other models. Therefore, improvement in the cross-lingual representations with the LEM strategy benefits the parallel data curation task as well.

Table 8 ChrF scores for the parallel data curation task. The scores have been reported on the Flores+ devtest. The values in brackets indicate the gains of XLM-R$_{LEM}$ compared to the XLM-R and the XLM-R$_{MLM+TLM}$ respectively. The best result for each language pair is highlighted in bold

	Sinhala - Tamil	English - Sinhala	English-Tamil
NLLB			
XLM-R	38.6	33.1	44.00
XLM-R$_{MLM+TLM}$	41.3	43.2	50.70
XLM-R$_{LEM}$	**(+3.5/+0.8) 42.1**	**(+10.8/+0.7) 43.9**	**(+7.2/+0.5) 51.2**
CCAligned			
XLM-R	37.2	10.2	5.2
XLM-R$_{MLM+TLM}$	42.3	33.9	31.5
XLM-R$_{LEM}$	**(+5.2/+0.3) 42.6**	**(+24.3/+0.6) 34.5**	**(+29.1/+2.8) 34.3**

Table 9 Scores for the sentiment analysis task in terms of Precision, Recall and F1. The results are reported on the English-Sinhala code-mixed evaluation set. The best result for each experiment is highlighted in bold

	Precision	Recall	F1
Fine-tuning on English-Sinhala Code-mixed Sentiment Analysis dataset			
XLM-R	88.72%	88.72%	88.72%
XLM-R$_{MLM+TLM}$	88.88%	88.88%	88.88%
XLM-R$_{LEM}$	**89.09%**	**89.33%**	**89.20%**
Fine-tuning on English Sentiment Analysis dataset (Zero-shot scores)			
XLM-R	74.19%	**82.14%**	**75.17%**
XLM-R$_{MLM+TLM}$	**74.74%**	79.57%	73.92%
XLM-R$_{LEM}$	69.18%	76.36%	68.13%
Fine-tuning on English-Sinhala Code-mixed Sentiment Analysis dataset			
XLM-R	88.72%	**91.00%**	89.77%
XLM-R$_{MLM+TLM}$	89.19%	90.83%	89.97%
XLM-R$_{LEM}$	90.63%	90.63%	**90.63%**

6.5 Code-mixed sentiment analysis

Table 9 summarizes the evaluation scores for the code-mixed sentiment analysis task on the English-Sinhala dataset. The XLM-R$_{LEM}$ model consistently outperformed both XLM-R and XLM-R$_{MLM+TLM}$ when fine-tuned directly on the code-mixed dataset.

In the zero-shot evaluation (fine-tuned on English-only data), all models showed reduced F1 scores. We suspect that the English monolingual fine-tuning step adversely impacted the cross-lingual alignment between the two languages.

When further fine-tuned on the code-mixed dataset, XLM-R$_{LEM}$ achieved the highest scores, indicating that the cross-lingual enhancement benefits the code-mixed sentiment classification task. These results highlight the promise of LEM in improving multilingual models, particularly for low-resource and code-mixed language applications.

Table 10 Ablation study results by changing the number of tokens masked in the linguistic entity. The results are for the Sinhala-Tamil language pair and the bitext mining downstream task. The best result corresponding to the FW, BW, IN, and Average columns is highlighted in bold

No. of Tokens Masked in Linguistic Entity	FW (Recall)	BW (Recall)	IN (Recall)	Average (Recall)
1	**92.13**	**93.18**	**89.10**	**91.47**
2	91.02	92.52	87.78	90.44
3	83.79	87.37	79.05	83.40
4	84.12	87.03	78.97	83.38

7 Ablation studies

7.1 The number of tokens for masking in LEM strategy

To evaluate the impact of the masked token count within linguistic entities, we conducted an ablation study by varying the number of masked tokens. We conducted this for the Sinhala-Tamil language pair. As reported in Table 16, the best result was returned for the 100% NE+15% MLM and 100% NE+15% TLM combinations in the LEM_{mono} and LEM_{para} steps, respectively. Therefore, we used this setting and the number of tokens for masking was varied. Results on the bitext mining task are reported in Table 10. It reveals a clear trend of decreasing performance.

When masked only one token per linguistic entity, the average performance across tasks was the highest. This outcome suggested that minimal masking preserved more contextual information, allowing the model to better capture dependencies critical for downstream tasks. As the masked token count increased to two or more, the average performance dropped. This drop was significant when increasing the token count to 3 and 4. This indicated that excessive masking had disrupted the contextual integrity of linguistic entities, which lead to suboptimal representations.

Interestingly, the performance drop became less pronounced when the number of masked tokens increased from 3 tokens to 4 tokens (a decrease of only 0.02). This suggested a potential saturation point where further masking within an entity had diminishing negative effects, as the model might already struggle to leverage the remaining context effectively.

7.2 Effect of noise in LEM Strategy

We investigated the impact of applying the LEM strategy to noisy data. This analysis provides critical insights into the robustness and adaptability of LEM when faced with real-world noisy data. We specifically focus on the English-Sinhala language pair and use the ParaCrawl[17] dataset.

As per Table 14, the current ablation revealed that the best results for English-Sinhala were achieved with the combination of 100%VB+15%MLM and 100%VB+15%TLM during the LEM_{mono} and LEM_{para} steps, respectively. We ran the LEM experiments with ParaCrawl data for the same combinations.

Table 11 presents the final scores. We observed that the results were comparable to those derived from the cleaner SiTa-Trilingual dataset. Further, in BW criteria the scores slightly

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

Table 11 Bitext mining results obtained using LEM-enhanced models on both high-quality and noisy web-crawled datasets. The best result corresponding to the FW, BW, IN, and Average columns is highlighted in bold

Dataset	Quality of the Dataset	FW (Recall)	BW (Recall)	IN (Recall)	Overall Average (Recall)
SiTa-Trilingual	High Quality	**95.09**	94.67	**93.42**	**94.39**
ParaCrawl	Noisy	**95.09**	**95.00**	92.49	94.19

surpass and in FW criteria the scores are the same as that obtained when using high-quality data. This equivalence underscores the resilience of the LEM strategy to noise in the training data.

The ability of LEM to maintain high performance in noisy settings highlights its practical applicability in low-resource scenarios, where parallel data is often noisy or inconsistent. This robustness not only complements our findings but also demonstrates that LEM can effectively mitigate the challenges associated with data quality, a common issue in low-resource language processing.

8 Discussion

The LEM strategy is very much driven by the accuracy of the underlying tools to identify the linguistic entities. The sub-optimal performance of the NER model and POS Taggers can affect the final results.

Although the NER model performs well with English sentences, we observe two main error types with it for Sinhala and Tamil language text, as shown in Table 17 in Appendix 1. As *False Positives*, we observe words which were not a part of the NE tagged as NEs. Secondly, in the category of *False Negatives*, the NER model fails to identify all the words belonging to the NE sequence, and incorrectly label these words as *Other* etc. Similar instances are found in PoS Tagging as well as per examples in Table 18 in Appendix 1 that resulted in *False Positives* and *False Negatives*. However for English language the returned PoS tags were mostly accurate.

As future work, we will continually pre-train a single multilingual model for multiple languages using the LEM strategy. This is in contrast to the current approach which yields specialized encoders for each language pair. Additionally, as a single multilingual model, its performance on downstream tasks can be analysed. Secondly, with the advancements in the field, we hope more sophisticated NER models and PoS Tagger tools might be introduced in future. We will re-evaluate the LEM performance upon the availability of such tools. Thirdly, we will investigate the impact of the LEM strategy on language families, examining its effectiveness across a broader linguistic spectrum.

9 Conclusion

Multilingual PLMs trained with masking strategies are less effective for downstream tasks. This research introduced LEM strategy to improve cross-lingual representations of existing multiPLMs. Here the objective is to mask a single token, specifically targeting linguistic entities (NEs, nouns and verbs). Extending this LEM strategy with parallel data yields even better results, as evidenced in low-resource language pairs such as English-Sinhala, English-Tamil, and Sinhala-Tamil. The improved cross-lingual representations showed superior performance on the three evaluation tasks.

Appendix A

Type of monolingual data parallel data for MLM

Table 12 shows the results corresponding to Fig. 4.

Table 12 Bitext mining recall scores for using pure monolingual data versus source and target sides from a parallel corpus (as monolingual data) for MLM experiments. Considering the average F (Forward), B (Backward), and I (Intersection) scores are highlighted in bold

Dataset	Dataset Size	Army			Hiru			ITN			Newsfirst			Averages		
		F	B	I	F	B	I	F	B	I	F	B	I	F	B	I
English - Sinhala																
SiTa	59333	88.33	91.00	85.33	92.03	93.36	89.70	91.67	92.67	88.67	91.67	95.33	90.00	**90.92**	**93.09**	**88.42**
MADLAD400	60000	82.67	88.33	78.00	85.05	91.36	82.00	85.33	86.00	79.67	91.67	92.67	87.67	86.18	89.59	81.83
MADLAD400	100000	86.67	91.67	83.33	91.69	96.01	91.03	88.00	90.33	83.00	91.33	95.00	89.00	89.42	93.25	86.59
English - Tamil																
SiTa	59333	84.00	86.00	77.67	80.33	75.00	68.33	81.56	82.21	78.52	90.67	91.00	87.00	**84.14**	**83.55**	**77.88**
MADLAD400	60000	81.67	78.67	69.33	75.33	69.67	60.67	81.18	77.15	69.77	90.00	86.67	81.67	82.05	78.04	70.36
MADLAD400	100000	81.33	79.67	71.67	77.67	71.33	62.67	78.86	76.17	68.79	88.67	88.00	82.00	81.63	78.79	71.28
Sinhala - Tamil																
SiTa	59333	86.75	88.08	81.46	88.00	89.33	84.00	93.33	92.67	89.33	90.33	94.00	89.00	**89.60**	**91.02**	**85.95**
MADLAD400	60000	84.77	89.73	80.46	86.00	89.00	83.00	92.67	92.00	89.00	89.00	92.67	85.67	88.11	90.85	84.53
MADLAD400	100000	84.11	88.08	78.81	86.00	89.33	81.33	90.67	93.67	87.33	88.67	92.33	85.00	87.36	90.85	83.12
MADLAD400	500000	82.12	83.11	75.17	85.67	88.33	79.67	87.67	91.00	83.67	87.67	90.67	82.33	85.78	88.28	80.21

Appendix B

Parallel data curation task NMT extrinsic evaluation results

The NMT evaluation scores for the parallel data curation task are reported in Table 13 for the Flores+ benchmark devtest evaluation set. While the discussion on the NMT results has been based using the ChrF++ metric, here we present the scores for the same experiments using NMT evaluation metrics sacreBLEU, multi-bleu, ChrF, ChrF++ and spBLEU.

Table 13 NMT scores on the Flores+ devtest using top 50,000 parallel sentences from the ranked NLLB and CCAligned corpus. For each language pair, the best score for each metric is highlighted in bold

	sacreBLEU	multi-bleu	ChrF	ChrF++	SpBLEU
NLLB					
Sinhala - Tamil					
XLM-R	2.6	2.58	38.6	33.58	11.9
XLM-R$_{MLM+TLM}$	3.2	3.23	41.3	35.99	14.5
XLM-R$_{LEM}$	**3.6**	**3.60**	**42.1**	**36.68**	**15.2**
English - Tamil					
XLM-R	6.2	6.18	44.00	38.28	18.40
XLM-R$_{MLM+TLM}$	9.2	9.16	50.70	45.35	25.20
XLM-R$_{LEM}$	**9.3**	**9.47**	**51.20**	**45.86**	**25.80**
English - Sinhala					
XLM-R	4.9	4.91	33.1	30.37	13.6
XLM-R$_{MLM+TLM}$	9.4	9.42	43.2	39.78	23.3
XLM-R$_{LEM}$	**9.9**	**9.85**	**43.9**	**40.31**	**23.8**
CCAligned					
Sinhala - Tamil					
XLM-R	2.2	2.23	37.2	32.43	10.6
XLM-R$_{MLM+TLM}$	3.7	3.74	42.3	36.02	15.2
XLM-R$_{LEM}$	**3.6**	**3.61**	**42.6**	**36.90**	**14.9**
English - Tamil					
XLM-R	0.2	0.17	5.2	5.80	1.2
XLM-R$_{MLM+TLM}$	3.2	3.24	31.5	28.55	11.5
XLM-R$_{LEM}$	**3.5**	**3.48**	**34.3**	**30.96**	**12.5**
English - Sinhala					
XLM-R	0.4	0.37	10.2	10.13	2.3
XLM-R$_{MLM+TLM}$	5.0	5.00	33.9	31.17	14.8
XLM-R$_{LEM}$	**5.1**	**5.09**	**34.5**	**31.71**	**15.3**

Appendix C

Linguistic entity masking ablation study

In this section, we present the full experiments along with the scores obtained during the ablation study of our LEM masking strategy. Tables 14, 15 and 16 contain the bitext mining recall scores for English-Sinhala, English-Tamil and Sinhala-Tamil language pairs, respectively.

Table 14 Ablation experiments and bitext mining scores for English-Sinhala language pair considering linguistic entity masking. The best result for each criterion is highlighted in bold

Experiment	Army			Hiru			ITN			Newsfirst			Averages		
	F	B	I	F	B	I	F	B	I	F	B	I	F	B	I
Baselines															
XLM-R	92.33	93.33	89.67	96.35	96.68	***95.68***	94.00	96.00	92.33	96.67	95.33	***94.33***	94.84	95.34	93.00
15%MLM	88.33	91.00	85.33	92.03	93.36	89.70	91.67	92.67	88.67	91.67	95.33	90.00	90.92	93.09	88.42
15% TLM on 15% MLM	91.33	92.67	88.67	94.35	95.68	93.36	94.00	94.00	90.67	94.67	95.00	92.67	93.59	94.34	91.34
LEM_{mono}															
100%NE+15% MLM	89.67	93.00	88.33	93.02	94.02	92.03	89.67	93.00	87.00	93.67	94.67	91.67	91.51	93.67	89.76
100% VB+15% MLM	89.67	93.33	87.33	94.02	95.02	92.69	92.00	93.67	89.67	93.00	95.33	92.33	92.17	94.34	90.51
100% NN+15% MLM	81.33	88.33	76.33	93.36	95.02	92.36	90.33	91.67	86.00	91.00	92.33	87.67	89.00	91.84	85.59
100% NE+ 100%VB+15% MLM	91.33	91.00	87.67	95.35	94.02	93.36	92.33	94.00	89.33	93.33	94.33	90.67	93.09	93.34	90.26
100% NE+ 100%NN+15% MLM	88.00	91.00	84.00	94.02	95.35	92.69	89.33	95.67	89.00	94.00	95.67	91.67	91.34	94.42	89.34
100% NE+ 100%VB+ 100%NN+15% MLM	89.67	92.33	87.00	94.02	94.02	91.69	92.33	95.00	91.00	94.00	92.33	90.33	92.50	93.42	90.01
MLM_{mono}+TLM_{para}															
100% NE+15% TLM on 15% MLM	90.00	91.67	87.33	95.02	95.35	93.36	94.00	***96.67***	92.67	96.67	96.67	93.33	93.92	95.09	91.67
100% VB+15% TLM on 15% MLM	91.67	90.33	86.67	94.35	95.02	92.69	93.00	95.33	89.67	95.00	94.67	91.67	93.50	93.84	90.17
100% NN+15% TLM on 15% MLM	89.00	92.00	85.00	93.36	95.02	91.36	94.33	96.00	92.33	94.67	95.00	92.00	92.84	94.50	90.17
100% NE+ 100%VB+15% TLM on 15% MLM	91.33	91.33	87.67	95.35	94.68	92.69	94.00	95.00	91.33	***97.33***	95.00	93.67	94.50	94.00	91.34
100% NE+ 100%NN+15% TLM on 15% MLM	88.67	91.00	85.00	94.35	95.35	93.02	94.00	96.00	92.00	93.67	95.00	91.33	92.67	94.34	90.34
100%NE+100%VB+100%NN+ 15%TLM on 15% MLM	90.67	91.33	87.33	94.68	***97.34***	94.35	93.67	95.00	91.00	94.33	96.33	92.33	93.34	95.00	91.25
15% TLM on (100%NE+15% MLM)	89.00	93.00	87.00	94.35	95.35	93.64	92.00	95.67	90.00	95.00	90.00	93.33	92.59	93.50	90.99
100% NE+15% TLM on (100%NE+15% MLM)	91.67	***95.33***	89.33	94.68	96.01	94.35	92.00	96.33	92.67	94.67	95.67	92.67	93.25	***95.84***	92.25
100% VB+15% TLM on (100%NE+15% MLM)	90.00	91.67	86.00	94.02	95.02	93.36	92.67	94.67	90.00	93.33	95.00	91.67	92.50	94.09	90.26

Table 14 continued

MLM_{mono}+TLM_{para}															
100% NN+15% TLM on (100%NE+15% MLM)	89.00	92.00	87.00	94.02	94.02	92.36	93.00	93.33	89.00	94.00	94.00	91.00	92.50	93.34	89.84
100% NE+ 100%VB+15% TLM on (100%NE+15% MLM)	89.67	93.33	88.00	95.02	94.68	93.36	92.00	95.33	90.00	95.67	95.33	93.33	93.09	94.67	91.17
100% NE+ 100%NN+15% TLM on (100%NE+15% MLM)	89.33	93.00	87.00	94.35	94.68	93.02	93.67	94.67	90.67	95.67	96.67	94.00	93.25	94.75	91.17
100%NE+100%VB+100%NN+15%TLM on (100%NE+15% MLM)	91.67	92.33	88.33	95.68	95.68	95.02	92.33	93.33	88.67	93.67	95.00	91.33	93.34	94.09	90.84
15% TLM on (100%VB+15% MLM)	91.67	92.00	89.00	94.35	96.01	94.02	94.33	95.00	91.67	95.67	96.00	93.33	94.00	94.75	92.00
100% NE+15% TLM on (100%VB+15% MLM)	90.33	91.67	87.67	95.02	96.35	94.35	93.67	94.33	90.00	96.67	95.67	93.67	93.92	94.50	91.42
100% VB+15% TLM on (100%VB+15% MLM)	91.67	93.33	90.67	***96.68***	95.35	95.35	95.33	94.33	93.67	96.67	95.67	94.00	***95.09***	94.67	***93.42***
100% NN+15% TLM on (100%VB+15% MLM)	88.33	91.67	86.00	95.02	95.02	93.36	93.00	93.67	89.67	92.67	94.67	91.33	92.25	93.75	90.09
100% NE+ 100%VB+15% TLM on (100%VB+15% MLM)	90.00	94.33	88.33	94.35	95.68	93.02	94.00	95.33	91.00	96.67	95.33	***94.33***	93.75	95.17	91.67
100% NE+ 100%NN+15% TLM on (100%VB+15% MLM)	89.67	91.33	86.33	94.68	95.68	93.69	93.67	94.33	91.33	95.67	96.33	93.67	93.42	94.42	91.26
100%NE+100%VB+100%NN+15%TLM on (100%VB+15% MLM)	92.00	92.33	87.33	95.35	95.68	94.02	93.00	94.00	89.67	95.67	95.00	93.00	94.00	94.25	91.00
15% TLM on (100%NN+15%MLM)	90.33	93.33	87.33	94.35	94.68	93.02	94.67	95.00	92.00	94.67	95.00	92.67	93.50	94.50	91.26
100% NE+15% TLM on (100%NN+15%MLM)	89.00	93.67	87.00	94.35	95.35	92.36	95.00	95.33	91.33	96.00	95.33	92.67	93.59	94.92	90.84
100% VB+15% TLM on (100%NN+15%MLM)	88.00	93.33	86.67	93.69	95.68	93.02	94.33	95.67	***94.67***	94.67	94.00	91.67	92.67	94.67	91.51
100% NN+15% TLM on (100%NN+15%MLM)	91.00	92.00	87.67	95.68	95.02	94.02	94.33	96.33	92.67	95.00	95.67	92.67	94.00	94.75	91.76

Table 14 continued

MLM_{mono}+TLM_{para}															
100% NE+ 100%VB+15% TLM on (100%NN+15%MLM)	90.67	93.67	87.67	95.02	94.68	93.02	95.00	95.67	92.33	94.33	94.00	91.67	93.75	94.50	91.17
100% NE+ 100%NN+15% TLM on (100%NN+15%MLM)	91.67	91.33	87.67	94.68	95.68	94.02	93.00	95.33	90.67	94.33	95.00	92.33	93.42	94.34	91.17
100%NE+100%VB+100%NN+15%TLM on (100%NN+15%MLM)	88.67	92.00	86.00	96.01	96.01	95.02	94.00	95.33	91.33	94.33	94.67	91.33	93.25	94.50	90.92
15% TLM on (100%NE+100%VB+15%MLM)	88.67	93.00	86.67	94.35	95.02	93.02	92.33	93.67	88.67	93.00	94.33	90.67	92.09	94.00	89.76
100% NE+15% TLM on (100%NE+100%VB+15%MLM)	89.67	91.33	87.00	94.68	95.68	93.69	93.67	94.33	93.67	95.67	94.33	91.33	93.42	93.92	91.42
100% VB+15% TLM on (100%NE+100%VB+15%MLM)	88.00	93.33	86.67	93.69	95.68	93.02	94.33	94.33	***94.67***	94.67	94.00	91.67	92.67	94.34	91.51
100% NN+15% TLM on (100%NE+100%VB+15%MLM)	88.67	93.00	86.67	93.67	95.02	93.02	94.00	93.67	90.00	93.00	94.33	90.67	92.33	94.00	90.09
100% NE+ 100%VB+15% TLM on (100%NE+100%VB+15%MLM)	91.33	92.67	89.00	94.68	95.68	93.69	94.00	94.67	91.67	95.33	95.33	93.00	93.84	94.59	91.84
100% NE+ 100%NN+15% TLM on (100%NE+100%VB+15%MLM)	91.00	91.33	87.67	94.02	94.68	92.36	94.67	94.67	91.67	95.67	95.33	93.00	93.84	94.00	91.17
100%NE+100%VB+100%NN+15%TLM on (100%NE+100%VB+15%MLM)	92.00	93.00	89.00	95.35	96.01	94.68	94.33	94.00	91.00	96.00	96.00	***94.33***	94.42	94.75	92.25
15% TLM on (100%NE+100%NN+15%MLM)	91.33	94.00	88.67	94.02	95.02	92.03	95.33	95.67	93.00	94.33	***97.67***	94.00	93.75	95.59	91.92
100% NE+15% TLM on (100%NE+100%NN+15%MLM)	87.67	90.33	***93.67***	94.02	95.35	93.02	***96.00***	94.67	92.67	93.67	94.67	91.67	92.84	93.75	92.76
100% VB+15% TLM on (100%NE+100%NN+15%MLM)	91.00	92.00	87.00	94.35	94.35	92.69	93.67	96.33	93.00	95.67	95.33	93.00	93.67	94.50	91.42
100% NN+15% TLM on (100%NE+100%NN+15%MLM)	88.33	91.67	84.33	95.02	95.35	93.64	94.67	94.67	91.67	94.33	95.33	92.33	93.09	94.25	90.49
100% NE+100%VB+15% TLM on (100%NE+100%NN+15%MLM)	90.00	93.33	86.67	94.68	94.68	93.36	93.33	93.00	89.33	96.33	96.00	***94.33***	93.59	94.25	90.92
100% NE+100%NN+15% TLM on (100%NE+100%NN+15%MLM)	87.67	90.33	84.00	95.68	96.01	94.35	92.00	95.00	90.33	94.67	96.33	93.00	92.50	94.42	90.42

Table 14 continued

MLM_{mono}+TLM_{para}															
100%NE+100%VB+100%NN+15%TLM on (100%NE+100%NN+15%MLM)	88.67	91.67	85.00	95.35	95.68	94.35	94.00	93.67	90.33	95.67	95.33	93.00	93.42	94.09	90.67
15% TLM on (100%NE+100%VB+100%NN+15%MLM)	***93.00***	91.67	88.00	95.35	96.01	94.02	94.67	95.00	92.00	93.67	94.67	91.67	94.17	94.34	91.42
100% NE+15% TLM on (100%NE+100%VB+100%NN+15%MLM)	89.00	91.00	84.67	95.35	96.35	94.68	***96.00***	95.33	93.00	96.00	95.67	93.33	94.09	94.59	91.42
100% VB+15% TLM on (100%NE+100%VB+100%NN+15%MLM)	89.67	92.67	87.00	95.35	95.35	93.67	95.00	93.33	91.67	95.67	93.33	91.00	93.92	93.67	90.83
100% NN+15% TLM on (100%NE+100%VB+100%NN+15%MLM)	89.67	91.33	86.00	95.02	95.35	93.33	93.67	94.67	91.33	93.67	94.00	90.33	93.00	93.84	90.25
100% NE+ 100%VB+15% TLM on (100%NE+100%VB+100%NN+15%MLM)	88.67	92.00	85.33	93.69	95.68	92.69	93.00	95.67	90.33	95.00	95.33	92.33	92.59	94.67	90.17
100% NE+ 100%NN+15% TLM on (100%NE+100%VB+100%NN+15%MLM)	86.67	91.67	84.67	96.35	96.01	94.68	94.33	95.67	92.33	94.33	94.67	91.33	92.92	94.50	90.75
100%NE+100%VB+100%NN+15%TLM on (100%NE+100%VB+100%NN+15%MLM)	91.33	92.00	88.00	95.68	95.68	94.02	93.67	94.00	91.33	93.67	94.00	90.33	93.59	93.92	90.92

Table 15 Ablation experiments and bitext mining scores for English-Tamil language pair considering linguistic entity masking . The best result for each criterion is highlighted in bold

Experiment	Army			Hiru			ITN			Newsfirst			Average		
	FW	BW	IN	FW	BW	IN	FW	BW	IN	FW	BW	IN	FW	BW	IN
Baselines															
XLM-R	86.67	88.33	82.00	83.00	78.33	72.67	83.22	83.56	78.86	92.33	91.33	89.33	86.31	85.39	80.71
15%MLM	84.00	86.00	77.67	80.33	75.00	68.33	81.56	82.21	78.52	90.67	91.00	87.00	84.14	83.55	77.88
15%TLM on 15%MLM	86.67	85.67	79.33	80.33	78.67	71.00	81.88	83.56	77.52	90.00	92.67	88.00	84.72	85.14	78.96
LEM$_{mono}$															
100% NE+15% MLM	86.00	86.67	81.00	79.33	75.33	66.67	81.21	81.21	74.83	93.00	92.00	90.00	84.89	83.80	78.12
100% VB+15% MLM	85.67	84.67	76.67	78.67	76.00	68.00	81.88	82.55	75.84	91.00	90.00	86.33	84.30	83.30	76.71
100% NN+15% MLM	83.33	84.67	77.00	73.67	72.67	61.67	75.84	82.22	70.13	90.00	91.00	87.00	80.71	82.64	73.95
100% NE+100%VB+15% MLM	83.00	86.67	77.67	77.67	74.33	65.00	81.21	83.56	75.84	89.00	88.67	84.00	82.72	83.31	75.63
100% NE+100%NN+15% MLM	82.67	85.33	75.00	75.33	72.67	62.00	80.54	84.23	74.48	90.33	90.00	86.33	82.22	83.06	74.45
100% NE+100% VB +100%NN+15% MLM	83.00	83.33	78.00	74.67	73.67	64.67	80.87	83.89	76.17	91.33	92.67	88.67	82.47	83.39	76.88
LEM$_{mono}$+LEM$_{para}$															
100% NE+15% TLM on 15%MLM	83.00	85.33	76.33	79.67	78.33	70.00	83.89	85.91	79.87	91.00	93.33	89.00	84.39	85.73	78.80
100% VB+15% TLM on 15%MLM	87.00	86.67	81.67	80.67	79.00	72.33	83.89	85.57	79.87	91.67	92.33	88.67	85.81	85.89	80.63
100% NN+15% TLM on 15%MLM	85.00	86.67	79.67	79.33	77.00	69.00	83.89	86.24	80.54	91.33	94.00	89.67	84.89	85.98	79.72
100% NE+100% VB+15% TLM on 15%MLM	85.67	76.67	69.33	79.67	76.67	69.33	83.22	84.23	77.85	92.00	92.00	89.33	85.14	82.39	76.46

Table 15 continued

$\mathbf{LEM}_{mono}$+$\mathbf{LEM}_{para}$															
100% NE+100% NN+15% TLM on 15%MLM	84.67	85.00	77.67	81.00	80.00	72.33	81.98	84.23	77.85	90.00	92.00	87.67	84.41	85.31	78.88
100% NE+100% VB+ 100% NN+ 15% TLM on 15%MLM	85.00	85.33	80.00	78.67	78.33	70.00	84.23	***88.59***	***80.87***	90.00	93.67	88.33	84.47	86.48	79.80
15% TLM on 100%NE+15%MLM	87.00	86.33	81.33	81.33	80.00	71.67	81.21	84.23	77.52	92.67	91.33	89.00	85.55	85.47	79.88
100%NE+15% TLM on 100%NE+15%MLM	87.67	87.00	81.67	82.00	***81.33***	***73.00***	81.88	84.23	77.18	91.33	92.67	88.33	85.72	86.31	80.05
100% VB+15% TLM on 100%NE+15%MLM	***88.33***	89.33	***83.67***	80.00	77.67	69.67	81.54	84.23	75.50	91.00	93.00	89.00	85.22	86.06	79.46
100% NN+15% TLM on 100%NE+15%MLM	86.33	87.67	80.33	81.33	79.67	70.67	80.54	84.56	76.51	92.00	91.33	88.00	85.05	85.81	78.88
100% NE+100% VB+15% TLM on 100%NE+15%MLM/	84.67	85.67	78.00	82.33	76.67	70.33	80.54	83.22	76.85	89.33	92.67	87.67	84.22	84.56	78.21
100% NE+100% NN+15% TLM on 100%NE+15%MLM/	84.67	85.67	78.00	82.33	76.67	70.33	80.54	83.22	76.85	89.33	92.67	87.67	84.22	84.56	78.21
100%NE+100%VB+100%NN+15%TLM on 100%NE+15%MLM/	85.00	84.33	78.33	78.00	76.67	67.33	79.53	83.89	75.84	92.33	92.00	90.00	83.72	84.22	77.88
15% TLM on (100%VB+15%MLM)	88.00	88.67	83.67	82.00	79.00	72.33	84.90	84.90	80.54	***93.33***	93.00	***91.00***	***87.06***	86.39	***81.88***
100% NE+ 15% TLM on (100%VB+15%MLM)	84.00	87.67	79.00	78.67	***81.33***	71.00	82.22	85.57	78.86	90.67	93.33	88.33	83.89	***86.98***	79.30
100% VB+ 15% TLM on (100%VB+15%MLM)	86.00	88.67	80.67	81.33	78.33	70.67	82.22	84.56	76.85	91.33	92.33	87.67	85.22	85.97	78.96
100% NN+15% TLM on (100%VB+15%MLM)	86.33	85.33	80.33	79.67	79.00	70.33	82.22	84.90	77.52	90.33	93.67	88.00	84.64	85.72	79.05

Table 15 continued

$\mathbf{LEM}_{mono}$+$\mathbf{LEM}_{para}$															
100% NE+ 100% VB+ 15% TLM on (100%VB+15%MLM)	85.67	88.00	80.33	80.33	76.00	69.00	81.54	83.58	77.18	90.67	93.00	88.00	84.55	85.14	78.63
100% NE+ 100% NN+ 15% TLM on (100%VB+15%MLM)	87.33	87.67	81.67	78.00	78.00	68.67	81.54	83.89	76.85	91.00	92.00	87.67	84.47	85.39	78.71
100% NE+ 100% NN+ 100%VB+ 15% TLM on (100%VB+15%MLM)	86.33	87.00	80.67	78.33	76.67	67.33	82.89	84.56	77.85	90.67	92.33	87.33	84.55	85.14	78.30
15% TLM on (100%NN+15%MLM)	84.67	88.33	81.00	81.00	77.33	69.67	83.22	85.91	78.86	91.67	92.33	89.67	85.14	85.98	79.80
100% NE+ 15% TLM on (100%NN+15%MLM)	85.33	86.67	79.00	78.33	76.33	67.33	81.88	84.56	76.85	91.00	91.33	87.67	84.14	84.72	77.71
100% VB+15% TLM on (100%NN+15%MLM)	84.67	87.67	80.67	78.67	76.00	67.33	79.19	84.29	75.50	90.00	92.67	88.00	83.13	85.16	77.88
100% NN+ 15% TLM on (100%NN+15%MLM)	85.00	87.00	79.33	77.33	76.33	66.00	80.87	83.56	74.83	89.67	92.67	87.00	83.22	84.89	76.79
100% NE+ 100% VB+ 15% TLM on (100%NN+15%MLM)	82.33	86.00	77.67	78.33	74.33	65.00	81.21	85.91	76.51	62.67	65.00	61.00	76.14	77.81	70.04
100% NE+ 100% NN+ 15% TLM on (100%NN+15%MLM)	86.00	87.00	80.00	78.00	76.67	66.67	79.53	84.23	74.16	88.67	92.33	86.67	83.05	85.06	76.87
100% NE+ 100% NN+ 100%VB+ 15% TLM on (100%NN+15%MLM)	86.33	***90.00***	82.00	76.00	78.33	66.33	79.87	85.91	76.16	90.33	92.00	87.67	83.13	86.56	78.04
15% TLM on (100%NE+100%VB+15%MLM)	85.33	89.33	80.00	80.00	75.33	67.33	85.34	84.29	78.86	90.00	91.67	87.00	85.17	85.16	78.30
100% NE+ 15% TLM on (100%NE+100%VB+15%MLM)	86.00	87.33	80.33	80.33	78.33	71.33	82.22	81.88	75.50	89.00	93.00	88.00	84.39	85.14	78.79

Table 15 continued

LEM$_{mono}$+LEM$_{para}$															
100% VB+15% TLM on (100%NE+100%VB+15%MLM)	84.67	87.67	79.33	78.00	75.33	67.00	83.89	84.56	77.85	88.33	92.00	86.67	83.72	84.89	77.71
100% NN+ 15% TLM on (100%NE+100%VB+15%MLM)	86.00	86.67	79.00	81.00	75.67	67.00	83.89	84.56	78.52	92.00	93.00	89.33	85.72	84.97	78.46
100% NE+100% VB+ 15% TLM on (100%NE+100%VB+15%MLM)	84.67	87.67	78.00	78.33	75.67	68.00	***90.87***	84.90	76.17	89.00	92.00	86.67	85.72	85.06	77.21
100% NE+ 100% NN+ 15% TLM on (100%NE+100%VB+15%MLM)	83.00	86.67	78.33	***84.67***	77.00	72.00	81.88	84.56	77.18	89.00	93.00	87.67	84.64	85.31	78.80
100%NE+100%VB+100%NN+15%TLM on (100%NE+100%VB+15%MLM)	87.67	86.33	80.00	79.00	78.00	69.33	82.89	84.29	78.52	90.00	93.67	88.67	84.89	85.57	79.13
15% TLM on (100%NE+100%NN+15%MLM)	86.00	89.33	80.00	80.00	76.33	69.67	85.34	85.24	78.86	90.00	92.67	88.00	85.33	85.89	79.13
100% NE+ 15% TLM on (100%NE+100%NN+15%MLM)	86.00	86.67	80.00	78.33	76.67	65.00	82.55	85.57	78.52	90.67	93.00	88.33	84.39	85.48	77.96
100% VB+15% TLM on (100%NE+100%NN+15%MLM)	84.67	87.67	79.33	80.67	78.67	70.00	81.54	85.23	78.86	89.33	93.00	87.00	84.05	86.14	78.80
100% NN+ 15% TLM on (100%NE+100%NN+15%MLM)	85.67	85.00	78.00	80.00	76.33	68.67	81.21	84.56	77.18	92.33	93.00	89.33	84.80	84.72	78.29
100% NE+100% VB+ 15% TLM on (100%NE+100%NN+15%MLM)	84.33	85.33	77.33	79.00	77.67	69.00	84.56	85.91	80.20	91.00	92.67	89.00	84.72	85.39	78.88
100% NE+100% NN+ 15% TLM on (100%NE+100%NN+15%MLM)	86.67	83.67	78.67	76.33	78.00	66.00	82.55	85.57	78.19	91.33	91.67	87.67	84.22	84.73	77.63
100%NE+100%VB+100%NN+15%TLM on (100%NE+100%NN+15%MLM)	82.33	86.00	76.67	77.00	79.00	68.67	81.21	82.55	75.84	91.00	91.33	88.00	82.89	84.72	77.29

Table 15 continued

LEM$_{mono}$+LEM$_{para}$															
15% TLM on (100%NE+100%VB+100%NN+15%MLM)	84.00	86.00	79.00	83.00	77.33	71.67	82.55	85.23	77.85	90.33	***94.33***	88.67	84.97	85.73	79.30
100% NE+ 15% TLM on (100%NE+100%VB+100%NN+15%MLM)	82.33	86.33	77.67	80.00	77.33	68.67	81.88	84.56	76.85	88.67	91.67	86.67	83.22	84.97	77.46
100% VB+15% TLM on (100%NE+100%VB+100%NN+15%MLM)	84.67	88.00	79.67	79.00	77.00	68.00	83.89	84.90	78.52	91.00	94.00	90.00	84.64	85.97	79.05
100% NN+ 15% TLM on (100%NE+100%VB+100%NN+15%MLM)	85.33	87.00	81.33	77.00	74.67	66.33	82.55	84.90	78.86	89.33	93.00	87.33	83.55	84.89	78.46
100% NE+100% VB+ 15% TLM on (100%NE+100%VB+100%NN+15%MLM)	82.67	85.67	78.33	76.67	75.00	66.00	83.22	85.91	77.52	88.00	92.67	85.67	82.64	84.81	76.88
100% NE+ 100% NN+ 15% TLM on (100%NE+100%VB+100%NN+15%MLM)	84.33	84.00	78.00	76.67	77.00	67.33	83.21	85.57	78.19	87.00	93.00	85.00	82.80	84.89	77.13
100%NE+100%VB+100%NN+15%TLM on (100%NE+100%VB+100%NN+15%MLM)	85.33	85.33	79.67	76.33	76.00	67.00	82.22	84.90	77.12	89.00	94.00	88.33	83.22	85.06	78.03

Table 16 Ablation experiments and bitext mining scores for Sinhala-Tamil language pair considering linguistic entity masking. The best result for each criterion is highlighted in bold

Experiment	Army			Hiru			ITN			Newsfirst			Average		
	FW	BW	IN	FW	BW	IN	FW	BW	IN	FW	BW	IN	FW	BW	IN
Baselines															
XLM-R	83.44	81.46	78.15	90.67	91.00	87.33	91.33	90.00	87.00	***93.67***	95.33	***92.33***	89.78	89.45	86.20
15%MLM	86.75	88.08	81.46	88.00	89.33	84.00	93.33	92.67	89.33	90.33	94.00	89.00	89.60	91.02	85.95
15%TLM on 15%MLM	87.75	90.40	83.11	88.67	93.33	86.33	93.00	94.33	90.00	91.33	94.33	89.67	90.19	93.10	87.28
LEM$_{mono}$															
100% NE+15% MLM	86.42	92.05	83.78	89.33	92.00	87.67	94.00	94.33	90.67	91.33	94.00	88.67	90.27	93.10	87.69
100% VB+15% MLM	83.44	88.08	78.81	87.33	90.33	83.33	92.33	94.00	88.00	90.00	92.00	87.67	88.28	91.10	84.45
100% NN+15% MLM	85.10	87.75	80.13	88.00	91.67	85.33	92.00	91.67	88.00	90.67	93.33	87.67	88.94	91.10	85.28
100% NE+100%VB+15% MLM	84.43	90.73	82.12	88.67	91.00	85.33	94.00	92.33	88.33	91.00	94.33	88.00	89.53	92.10	85.95
100% NE+100%NN+15% MLM	85.43	88.08	79.47	88.33	89.67	85.00	95.00	94.67	91.33	92.33	93.33	89.67	90.27	91.44	86.37
100% NE+100% VB +100%NN+15% MLM	83.11	88.41	79.80	86.67	91.33	84.33	91.67	89.67	85.33	90.33	93.67	88.33	87.94	90.77	84.45
LEM$_{mono}$**+LEM**$_{para}$															
100% NE+15% TLM on 15%MLM	89.07	90.73	85.10	89.33	91.00	85.67	95.67	94.67	92.67	91.00	93.33	88.33	91.27	92.43	87.94
100% VB+15% TLM on 15%MLM	88.41	91.00	84.77	87.67	91.00	85.67	93.67	93.67	90.67	92.33	93.00	90.00	90.52	92.17	87.78
100% NN+15% TLM on 15%MLM	88.74	90.07	84.44	89.67	91.67	86.67	94.67	93.33	90.67	92.00	91.67	87.33	91.27	91.68	87.28
100% NE+100% VB+15% TLM on 15%MLM	86.75	90.73	83.11	89.67	90.33	86.33	92.67	92.67	88.67	92.33	***95.33***	90.67	90.36	92.27	87.19
100% NE+100% NN+15% TLM on 15%MLM	86.42	90.73	83.11	87.33	89.33	84.00	94.67	93.33	91.67	92.00	93.67	88.33	90.11	91.77	86.78
100% NE+100% VB+ 100% NN+ 15% TLM on 15%MLM	85.43	91.39	81.79	89.00	92.33	86.67	93.67	93.33	88.67	90.33	93.00	87.00	89.61	92.51	86.03
15% TLM on 100%NE+15%MLM	87.09	89.73	83.44	89.33	92.00	86.33	94.33	92.67	89.33	92.00	93.67	89.67	90.69	92.02	87.19
15% NE+15%TLM on 100%NE+15%MLM	88.33	***93.33***	87.33	88.33	93.33	87.33	93.33	94.00	89.00	92.00	93.67	89.67	90.50	***93.58***	88.33
100% VB+15% TLM on 100%NE+15%MLM	86.42	90.07	83.11	90.00	92.00	87.67	94.33	93.00	90.33	92.00	94.67	90.00	90.69	92.43	87.78
100% NN+15% TLM on 100%NE+15%MLM	86.09	91.72	83.78	89.67	92.67	87.67	95.33	95.00	91.67	92.67	93.33	89.67	90.94	93.18	88.19

Table 16 continued

Experiment	Army			Hiru			ITN			Newsfirst			Average		
	FW	BW	IN	FW	BW	IN	FW	BW	IN	FW	BW	IN	FW	BW	IN
100% NE+100% VB+15% TLM on 100%NE+15%MLM/	87.09	90.07	84.11	89.00	91.33	86.67	95.67	94.67	91.67	90.67	92.33	88.67	90.61	92.10	87.78
100% NE+100% NN+15% TLM on 100%NE+15%MLM/	86.09	90.73	83.11	90.00	***93.67***	***89.33***	94.33	94.00	90.33	91.00	***95.33***	89.33	90.36	93.43	88.03
100%NE+100%VB+100%NN+15%TLM on 100%NE+15%MLM/	86.09	93.05	84.11	89.67	92.00	88.00	95.67	95.00	***93.33***	91.00	94.00	89.33	90.61	93.51	88.69
15% TLM on (100%VB+15%MLM)	89.07	88.41	83.44	89.67	92.33	87.00	93.33	93.67	90.00	91.00	91.67	87.33	90.77	91.52	86.94
100% NE+ 15% TLM on (100%VB+15%MLM)	87.75	88.74	83.11	90.00	91.00	86.33	95.00	94.67	91.67	93.00	91.67	88.00	91.44	91.52	87.28
100% VB+ 15% TLM on (100%VB+15%MLM)	88.74	90.75	84.44	89.67	92.00	86.67	92.67	94.33	89.33	92.33	93.33	89.33	90.85	92.60	87.44
100% NN+15% TLM on (100%VB+15%MLM)	86.42	90.73	84.11	89.67	92.33	86.33	91.33	93.33	88.00	92.00	94.00	89.67	89.86	92.60	87.03
100% NE+ 100% VB+ 15% TLM on (100%VB+15%MLM)	85.76	88.08	81.13	90.33	92.33	88.00	93.00	91.33	88.67	92.33	92.00	88.33	90.36	90.94	86.53
100% NE+ 100% NN+ 15% TLM on (100%VB+15%MLM)	87.09	89.07	82.78	88.67	92.33	85.00	93.33	93.67	89.67	92.00	91.33	87.00	90.27	91.60	86.11
100% NE+ 100% NN+ 100%VB+ 15% TLM on (100%VB+15%MLM)	85.77	90.40	83.11	89.67	92.00	86.00	92.33	93.67	88.67	92.00	92.00	88.00	89.94	92.02	86.44
15% TLM on (100%NN+15%MLM)	88.41	91.39	85.76	88.33	92.67	85.67	95.67	***95.67***	91.67	91.00	93.67	89.33	90.85	93.35	88.11
100% NE+ 15% TLM on (100%NN+15%MLM)	***89.40***	92.72	87.42	90.13	93.33	88.67	96.67	93.00	90.67	92.33	93.67	89.67	***92.13***	93.18	89.10
100% VB+15% TLM on (100%NN+15%MLM)	87.75	90.07	83.11	87.67	92.00	85.00	93.33	93.33	89.00	89.67	92.33	87.67	89.60	91.93	86.19
100% NN+ 15% TLM on (100%NN+15%MLM)	85.43	90.73	82.12	88.67	93.00	86.67	95.33	93.67	91.00	93.00	92.67	89.33	90.61	92.52	87.28
100% NE+ 100% VB+ 15% TLM on (100%NN+15%MLM)	86.09	90.40	82.78	***91.33***	92.00	88.33	94.67	94.67	91.33	91.33	93.00	88.33	90.86	92.52	87.69
100% NE+ 100% NN+ 15% TLM on (100%NN+15%MLM)	87.75	89.40	83.11	88.33	92.00	85.67	95.00	93.67	90.67	92.00	93.00	89.33	90.77	92.02	87.19
100% NE+ 100% NN+ 100%VB+ 15% TLM on (100%NN+15%MLM)	85.43	92.05	83.78	89.33	93.00	87.00	94.33	93.33	90.00	90.67	92.67	88.00	89.94	92.76	87.19

Table 16 continued

Experiment	Army			Hiru			ITN			Newsfirst			Average		
	FW	BW	IN	FW	BW	IN	FW	BW	IN	FW	BW	IN	FW	BW	IN
15% TLM on (100%NE+100%VB+15%MLM)	86.09	91.06	83.44	90.33	90.67	87.33	96.33	94.33	92.33	92.33	94.33	90.00	91.27	92.60	88.28
100% NE+ 15% TLM on (100%NE+100%VB+15%MLM)	86.75	90.07	83.44	89.33	91.33	86.67	93.67	94.67	91.33	92.00	93.33	89.67	90.44	92.35	87.78
100% VB+15% TLM on (100%NE+100%VB+15%MLM)	84.44	91.39	81.46	89.33	91.67	85.00	95.33	93.33	91.00	92.00	93.33	89.33	90.28	92.43	86.70
100% NN+ 15% TLM on (100%NE+100%VB+15%MLM)	85.76	92.05	85.00	90.67	93.00	88.33	95.67	95.33	92.67	89.33	92.67	86.33	90.36	93.26	88.08
100% NE+100% VB+ 15% TLM on (100%NE+100%VB+15%MLM)	87.09	89.73	83.11	90.67	92.33	88.33	94.67	92.67	89.33	92.00	93.00	90.00	91.10	91.93	87.69
100% NE+ 100% NN+ 15% TLM on (100%NE+100%VB+15%MLM)	85.76	92.05	85.00	90.67	91.67	87.67	95.67	94.00	91.33	90.67	93.67	88.33	90.69	92.85	88.08
100%NE+100%VB+100%NN+15%TLM on (100%NE+100%VB+15%MLM)	86.75	92.05	84.77	88.33	90.00	86.67	94.33	94.67	90.67	89.67	93.67	87.67	89.77	92.60	87.44
15% TLM on (100%NE+100%NN+15%MLM)	86.75	91.72	83.11	90.67	92.33	88.67	95.67	93.67	91.33	92.00	94.33	90.00	91.27	93.01	88.28
100% NE+ 15% TLM on (100%NE+100%NN+15%MLM)	86.75	87.47	81.79	90.33	93.00	88.00	95.33	93.33	90.33	***93.67***	94.00	90.33	91.52	91.95	87.61
100% VB+15% TLM on (100%NE+100%NN+15%MLM)	87.75	90.40	83.44	89.33	92.33	86.67	95.00	94.00	91.33	91.67	94.67	89.67	90.94	92.85	87.78
100% NN+ 15% TLM on (100%NE+100%NN+15%MLM)	87.75	90.40	84.77	87.67	89.67	84.00	95.33	95.33	92.00	90.00	93.67	88.67	90.19	92.27	87.36
100% NE+100% VB+ 15% TLM on (100%NE+100%NN+15%MLM)	85.76	87.75	81.13	90.33	90.67	86.00	94.33	92.00	89.00	92.33	93.67	90.00	90.69	91.02	86.53
100% NE+100% NN+ 15% TLM on (100%NE+100%NN+15%MLM)	87.75	90.07	84.44	90.33	92.00	86.67	96.00	94.00	91.67	93.00	93.67	89.00	91.77	92.43	87.94
100%NE+100%VB+100%NN+15%TLM on (100%NE+100%NN+15%MLM)	85.76	88.74	***91.46***	89.67	92.00	86.00	94.33	93.67	91.00	93.00	94.00	89.67	90.69	92.10	***89.53***

Table 16 continued

Experiment	Army			Hiru			ITN			Newsfirst			Average		
	FW	BW	IN	FW	BW	IN	FW	BW	IN	FW	BW	IN	FW	BW	IN
15% TLM on (100%NE+100%VB+100%NN+15%MLM)	86.09	89.40	82.45	89.33	92.00	87.00	94.33	91.67	88.33	91.33	92.00	86.33	90.27	91.27	86.03
100% NE+ 15% TLM on (100%NE+100%VB+100%NN+15%MLM)	88.76	89.40	84.44	89.67	91.33	87.33	95.00	94.00	90.67	90.33	92.00	87.00	90.94	91.68	87.36
100% VB+15% TLM on (100%NE+100%VB+100%NN+15%MLM)	85.76	89.40	82.12	90.33	93.33	88.67	94.67	94.00	90.67	93.00	93.67	90.00	90.94	92.60	87.86
100% NN+ 15% TLM on (100%NE+100%VB+100%NN+15%MLM)	85.43	90.73	82.78	89.67	91.67	87.33	95.33	92.33	90.33	90.00	93.33	86.67	90.11	92.02	86.78
100% NE+100% VB+ 15% TLM on (100%NE+100%VB+100%NN+15%MLM)	86.42	91.00	82.78	88.67	91.67	86.33	94.00	92.67	89.33	93.00	93.67	90.33	90.52	92.25	87.19
100% NE+ 100% NN+ 15% TLM on (100%NE+100%VB+100%NN+15%MLM)	86.75	90.73	84.11	90.67	93.33	88.33	***97.33***	92.67	92.33	91.33	92.00	86.67	91.52	92.18	87.86
100%NE+100%VB+100%NN+15%TLM on (100%NE+100%VB+100%NN+15%MLM)	87.42	89.40	82.78	89.33	92.33	86.33	94.67	95.00	91.33	91.33	94.00	88.67	90.69	92.68	87.28

Appendix D

Limitations of the NER model and Pos tagger

Examples highlighting the error categories found with the NER model and PoS taggers (Sect. 8) are shown in Table 17 and Table 18 respectively.

Table 17 Examples of incorrect identification and labeling of NEs. We identify mainly two error categories: false positives and false negatives, where the NER model underperforms

NER Labelling	
False Positives: Incorrect words tagged as NEs	
Ta	அரசாங்க B-ORG அபிவிருத்தி O முன்னெடுப்புக்களாக O நடைமுறைப்படுத்தப்பட O வேண்டிய O அபிவிருத்திக் B-MISC செயற்றிட்டங்கள் O மற்றும் O நிகழ்ச்சித் O திட்டங்களுக்கு B-MISC அவசியமான O நிதி O வசதிகளை O வழங்குவதற்கு O நிரல் O அமைச்சுக்களுடனும் O மற்றும் O அபிவிருத்திப் O பங்காளர்களுடனும் O ஒருங்கிணைப்பு O நடவடிக்கைகளை O மேற்கொள்ளல். O **அரசாங்க B-ORG (Government) is not a NE therefore tagged as O**
False Negatives: NEs, not identified during NER	
Si	අපනයන O සංවර්ධන I-MISC සහ O උපදේශන O සේවා O **The entire sentence should be NE therefore the correct tag sequence should be අපනයන B-ORG සංවර්ධන I-ORG සහ I-ORG උපදේශන I-ORG සේවා I-ORG (Export Development and Consultancy Services)**
Si	ඩී. B-PER ඩී. I-PER විමලරත්න I-PER මයා O **මයා O (Mr.) should be I-PER**
Ta	தலைமை O உரையானது O ஸ்ரீ B-PER ஜயவர்தனபுர I-PER பல்கலைக்கழக I-MISC துணைவேந்தர் B-MISC பேராசிரியர் B-MISC சம்பத் B-PER அமரதுங்க I-PER அவர்களால் O ஆற்றப்படும். O **The origanisation ஸ்ரீ B-PER ஜயவர்தனபுர I-PER பல்கலைக்கழக I-MISC (Sri Jayawardanapura Univerity) should be identified as a single NE and the correct tag sequence is ஸ்ரீ B-ORG ஜயவர்தனபுர I-ORG பல்கலைக்கழக I-ORG**
Ta	திரு. O எச். B-PER எஸ். I-PER எல். I-PER ராஜபக்ஸ I-PER திருமதி. O பீ. I-PER கே.எஸ்.எம். I-PER சியாமா I-PER சமரவீர I-PER **Both salutations திரு. O (Mr.) and திருமதி. O (Mrs.) should be B-PER .** **Hence எச். B-PER should be கே.எஸ்.எம். I-PER**

Table 18 Examples of incorrect identification and labeling of PoS Tags. We identify mainly two error categories: false positives and false negatives, where the Pos Tagger underperforms

PoS Tagging	
False Positives: Nouns/Verbs incorrectly identified during PoS Tagging	
Ta	எனவே, NOUN 2019 NUM வரவு NOUN செலவுத்திட்ட NOUN தரவுகளை NOUN இந்த DET இணைய NOUN முறைமையில் NOUN உட்படுத்துவது NOUN கட்டாயமானதாகும் VERB **In the sentence எனவே, NOUN (Therefore) should be ADVERB, உட்படுத்துவது NOUN (subject to) should be VERB and கட்டாயமானதாகும் VERB (is compulsory) should be an ADJ.**
False Negatives: Nouns/Verbs not identified during PoS Tagging	
Si	මැද JJ ඇළ NNC පුනරුත්ථාපනය NNC කිරීම NNC **කිරීම NNC (do) should be a VERB**
Si	10. NUM පොලොන්නරුව NNC මහාධිකරණ NNJ විනිසුරු NNC නිල JJ J නිවස ??? ඉදි RPCV කිරීම NNC **නිවස ??? (house) should be a NOUN**
Ta	கொள்வனவு NOUN செய்யப்பட்ட VERB நூல்கள் NOUN அறநெறிப் NOUN பாடசாலை NOUN மாணவர்களுக்கு NOUN விநியோகிக்கப்பட்டன NONE **In the sentence கொள்வனவு NOUN (Purchased) should be a VERB, அறநெறிப் NOUN (Moral) should be ADJ and விநியோகிக்கப்பட்டன NONE (were distributed) should be a VERB**

Acknowledgements I would like to thank and acknowledge the National Language Processing (NLP) Centre at the University of Moratuwa for providing the GPUs to execute the experiments related to the research.

Author Contributions Aloka Fernando contributed to the study conception. Both Aloka Fernando and Surangika Ranathunga decided on the steps to execute in this research work. Aloka Fernando conducted the literature review, material preparation, implementation, and execution of experiments. Aloka Fernando wrote the manuscript's first draft, and Surangika Ranathunga did the reviewing and final update. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This research was funded by the Google Award for Inclusion Research (AIR) 2022, received by Dr Surangika Ranathunga and Dr Nisansa de Silva.

Availability of data and materials We have only used publicly available data in this research. Where applicable, the citations and/or URLs were provided under the relevant sections.

Declarations

Competing interests The authors declare no competing interests.

References

1. 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). p. 4171–4186
2. 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 8440–8451
3. Ács J, Lévai D, Kornai A (2021) Evaluating Transferability of BERT Models on Uralic Languages. In: Proceedings of the seventh international workshop on computational linguistics of Uralic Languages p. 8–17
4. Dhananjaya V, Demotte P, Ranathunga S, Jayasena S (2022) BERTifying Sinhala-A comprehensive analysis of pre-trained language models for sinhala text classification. In: Proceedings of the thirteenth language resources and evaluation conference 7377–7385
5. Hu J, Ruder S, Siddhant A, Neubig G, Firat O, Johnson M (2020) XTREME: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In: Proceedings of the 37th international conference on machine learning p. 4411–4421
6. Hu J, Johnson M, Firat O, Siddhant A, Neubig G (2021) Explicit Alignment Objectives for Multilingual Bidirectional Encoders. In: Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies p. 3633–3643
7. Conneau A, Lample G (2019) Cross-lingual language model pretraining. Adva Neural Inf proc Syst. 32
8. Nastase V, Merlo P (2024) Tracking linguistic information in transformer-based sentence embeddings through targeted sparsification. In: Proceedings of the 9th workshop on representation learning for NLP (RepL4NLP-2024) p. 203–214
9. Nastase V, Merlo P (2023) Grammatical information in BERT sentence embeddings as two-dimensional arrays. In: Proceedings of the 8th workshop on representation learning for NLP (RepL4NLP 2023); 22–39

10. Aoyama T, Schneider N (2022) Probe-less probing of BERT's layer-wise linguistic knowledge with masked word prediction. In: Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies: student research workshop; 195–201
11. Sun Y, Wang S, Li Y, Feng S, Chen X, Zhang H, et al (2019) Ernie: Enhanced representation through knowledge integration. arXiv preprint arXiv:1904.09223. 2019
12. Joshi M, Chen D, Liu Y, Weld DS, Zettlemoyer L, Levy O (2020) SpanBERT: improving pre-training by representing and predicting spans. Trans Assoc Comput Linguist 8:64–77
13. Levine Y, Lenz B, Lieber O, Abend O, Leyton-Brown K, Tennenholtz M, et al (2020) PMI-Masking: Principled masking of correlated spans. In: International conference on learning representations
14. Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) SQuAD: 100,000+ Questions for machine comprehension of text. In: Proceedings of the 2016 conference on empirical methods in natural language processing; 2383–2392
15. Lai G, Xie Q, Liu H, Yang Y, Hovy E (2017) RACE: Large-scale ReAding comprehension dataset from examinations. In: Proceedings of the 2017 conference on empirical methods in natural language processing 785–794
16. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S (2018) GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP 353–355
17. Zhuang Y (2023) Heuristic masking for text representation pretraining. In: ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE p. 1–5
18. Golchin S, Surdeanu M, Tavabi N, Kiapour A (2023) Do not Mask Randomly: Effective Domain-adaptive Pre-training by Masking In-domain Keywords. In: Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023) 13–21
19. Wettig A, Gao T, Zhong Z, Chen D (2023) Should you mask 15% in masked language modeling? In: Proceedings of the 17th conference of the european chapter of the association for computational linguistics; 2023. p. 2977–2992
20. Feng F, Yang Y, Cer D, Arivazhagan N, Wang W (2022) Language-agnostic BERT Sentence Embedding. In: Proceedings of the 60th annual meeting of the association for computational linguistics (Volume 1: Long Papers); 2022. p. 878–891
21. Ranathunga S, Ranasinghea A, Shamala J, Dandeniyaa A, Galappaththia R, Samaraweeraa M (2024) A Multi-way Parallel Named Entity Annotated Corpus for English, Tamil and Sinhala. arXiv preprint arXiv:2412.02056
22. Akbik A, Blythe D, Vollgraf R (2018) Contextual String Embeddings for Sequence Labeling. In: COLING 2018, 27th International conference on computational linguistics 1638–1649
23. Fernando S, Ranathunga S (2018) Evaluation of different classifiers for sinhala pos tagging. In, (2018) Moratuwa Eng Res Conf (MERCon). IEEE 96–101
24. Fernando S, Ranathunga S, Jayasena S, Dias G (2016) Comprehensive part-of-speech tag set and svm based pos tagger for sinhala. In: Proceedings of the 6th workshop on south and southeast asian natural language processing (WSSANLP2016) 173–182
25. Sarveswaran K, Dias G (2020) ThamizhiUDp: A dependency parser for tamil. In: Proceedings of the 17th international conference on natural language processing (ICON) 200–207
26. 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 3197–3203
27. Fernando A, Ranathunga S, Sachintha D, Piyarathna L, Rajitha C (2023) Exploiting bilingual lexicons to improve multilingual embedding-based document and sentence alignment for low-resource languages. Knowled Inf Syst 65(2):571–612
28. Artetxe M, Schwenk H (2019) Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Trans Assoc Comput Linguist 7:597–610
29. Yang Y, Cer D, Ahmad A, Guo M, Law J, Constant N, et al (2020) Multilingual universal sentence encoder for semantic retrieval. In: Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations 87–94
30. Schwenk H, Wenzek G, Edunov S, Grave É, 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) 6490–6500
31. Costa-jussà MR, Cross J, Çelebi O, Elbayad M, Heafield K, Heffernan K, et al (2022) No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672

32. 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. https://doi.org/10.1162/tacl_a_00447
33. Ranathunga S, De Silva N, Menan V, Fernando A, Rathnayake C (2024) Quality does matter: a detailed look at the quality and utility of web-mined parallel corpora. In: Graham Y, Purver M, editors. Proceedings of the 18th conference of the european chapter of the association for computational linguistics (Volume 1: Long Papers). St. Julian's, Malta: association for computational linguistics; 2024. p. 860–880. Available from: https://aclanthology.org/2024.eacl-long.52
34. Post M (2018) A Call for Clarity in Reporting BLEU Scores. In: Proceedings of the third conference on machine translation: research papers. belgium, brussels: association for computational linguistics 186–191. Available from: https://www.aclweb.org/anthology/W18-6319
35. Popović M (2015) chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the tenth workshop on statistical machine translation 392–395
36. Popović M (2017) chrF++: words helping character n-grams. In: Proceedings of the second conference on machine translation 612–618
37. Goyal N, Gao C, Chaudhary V, Chen PJ, Wenzek G, Ju D et al (2022) The flores-101 evaluation benchmark for low-resource and multilingual machine translation. Trans Assoc Comput Linguist 10:522–538
38. Barbieri F, Anke LE, Camacho-Collados J (2022) XLM-T: Multilingual language models in twitter for sentiment analysis and beyond. In: Proceedings of the thirteenth language resources and evaluation conference 258–266
39. Myers-Scotton C, Jake J Duelling languages. Grammatical structure in Codeswitching–Clarendon Press. Oxford
40. 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 6282–6293
41. Ranathunga S, de Silva N (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 823–848
42. de Silva N (2023) Survey on publicly available sinhala natural language processing tools and research. arXiv preprint arXiv:1906.02358v20
43. Kudugunta S, Caswell I, Zhang B, Garcia X, Xin D, Kusupati A, et al (2024) Madlad-400: A multilingual and document-level large audited dataset. Adv Neural Inf Proc Syst 36
44. 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
45. El-Kishky A, Chaudhary V, Guzmán F, Koehn P (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) 5960–5969
46. 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 4555–4567
47. Rathnayake H, Sumanapala J, Rukshani R, Ranathunga S (2022) Adapter-based fine-tuning of pre-trained multilingual language models for code-mixed and code-switched text classification. Knowled Inf Syst 64(7):1937–1966
48. Koloski B, Škrlj B, Robnik-Šikonja M, Pollak S (2023) Measuring catastrophic forgetting in cross-lingual transfer paradigms: exploring tuning strategies. arXiv preprint arXiv:2309.06089
49. Udawatta P, Udayangana I, Gamage C, Shekhar R, Ranathunga S (2024) Use of prompt-based learning for code-mixed and code-switched text classification. World Wide Web 27(5):63
50. 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 48–53
51. Kocmi T, Zouhar V, Federmann C, Post M (2024) Navigating the metrics maze: reconciling score magnitudes and accuracies. In: Ku LW, Martins A, Srikumar V, editors. Proceedings of the 62nd annual meeting of the association for computational linguistics (Volume 1: Long Papers). Bangkok, Thailand: Association for Computational Linguistics 1999–2014. Available from: https://aclanthology.org/2024.acl-long.110

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 has 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. (PDF) Exploiting bilingual lexicons to improve multilingual embedding-based document and sentence alignment for low-resource languages.

Dr Surangika Ranathunga is a Senior Lecturer in the School of Mathematical and Computational Sciences, Massey University, New Zealand. She holds a PhD from the University of Otago, New Zealand. Her research interests include Natural Language Processing and Machine Learning. She is particularly interested in low-resource language computing. She was a recipient of the Google Award for Inclusion Research, 2022.