# DATA AUGMENTATION TO INDUCE HIGH QUALITY PARALLEL DATA FOR LOW-RESOURCE NEURAL MACHINE TRANSLATION

W.A.S.A Fernando

208035D

Doctor of Philosophy

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

September 2025

# DATA AUGMENTATION TO INDUCE HIGH QUALITY PARALLEL DATA FOR LOW-RESOURCE NEURAL MACHINE TRANSLATION

W.A.S.A Fernando

208035D

Dissertation submitted in partial fulfillment of the requirements for the degree
Doctor of Philosophy

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

September 2025

## DECLARATION

I declare that this is my own work and this Dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 03.09.2025

The supervisors should certify the Dissertation with the following declaration.

The above candidate has carried out research for the Doctor of Philosophy Dissertation under our supervision. We confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr. Nisansa de Silva

Signature of the Supervisor: Digitally signed by Nisansa de Silva Date: 2025.09.03 16:15:47 +05'30' Date: 03.09.2025

Name of Supervisor: Dr. Surangika Ranathunga

Signature of the Supervisor: Surangika Ranathunga Digitally signed by Surangika Ranathunga Date: 2025.09.03 22:07:46 +12'00' Date: 03.09.2025

**DEDICATION**

To *The God*

For the blessings throughout my life and for aligning
this opportunity for me.

To My *Thaththa*, Ranjith and *Amma*, Indrani,

For your selfless love and endless sacrifices for shaping me into the
person I am today.

To My *Husband*, Kumudu and
our *Children*, Kaviru, Adeesha and Asheth,

For your endless love, encouragement and strength you have
always given me.

## ACKNOWLEDGEMENT

I am grateful for my supervisors, Dr Nisansa de Silva and Dr Surangika Ranathunga, for the unreserved guidance and support provided during the course of my PhD journey. I highly appreciate Dr. Surangika Ranathunga for inspiring me to take this path and for continuing to mentor me with dedication, even after her transition to another University. I extend the same appreciation to Dr. Nisansa de Silva, for his willingness be my supervisor and for the invaluable mentorship since then. I consider myself truly fortunate to have had the opportunity to work under their supervision.

I would like to sincerely thank Prof. Gihan Dias for giving me the opportunity to join the National Languages Processing Centre (NLPC) at the University of Moratuwa as a Research Engineer. This opportunity laid the foundation for my research. I am also grateful for his encouragement and unwavering support throughout this academic journey.

I wish to thank Dr Uthayasanker Thayasivam, current Head of the Department of Computer Science and Engineering for his unresearved support extended during this time, and for his continuous motivation towards completing the PhD.

I am thankful to Prof. Sanath Jayasena and Dr. Kutila Gunasekara, CSE Research Coordinator, for their involvement and helpful contributions during my academic journey.

I am grateful to my progress panel, Prof. Asoka Karunananda, Dr. Charith Chitraranjan, and Dr. Lochandaka Ranathunga, for their valuable feedback and insightful suggestions that helped enrich and strengthen my research. I am thankful to all the lecturers at the Department of Computer Science and Engineering for extending their support throughout this time.

Further, I would like to acknowledge the Accelerating Higher Education Expansion and Development (AHEAD) Operation of the Ministry of Education, funded by the World Bank for facilitating the initial part of the research. Secondly, I wish to acknowledge the Senate Research Committee (SRC) grant of the University of Moratuwa, Sri Lanka, for funding the second part of the research. I wish to acknowledge the LK domain registry for funding me with the Prof. V. K. Samaranayake top-up grant during the third phase of this research. The final phase of this research was funded by the Google Award for Inclusion Research (AIR) 2022 received by Dr Surangika Ranathunga and Dr Nisansa de Silva. I would like to thank and acknowledge the National Languages Processing (NLP) Centre, at the University of Moratuwa for providing the GPUs to execute the experiments related to the research.

## ABSTRACT

Supervised Neural Machine Translation (NMT) models rely on parallel data to produce reliable translations. A parallel dataset refers to a collection of texts in two or more languages in which each sentence in one language is aligned with its corresponding translation counterpart in the other language. NMT models have produced state-of-the-art results for High-Resource Languages. HRLs refer to languages that have linguistic resources and tool support. In Low-Resource settings, which means for languages with limited or no linguistic resources and/or tools, NMT performance is suboptimal due to two challenges: the parallel data scarcity problem and the presence of noise in the available parallel datasets. Data augmentation (DA) is a viable approach to address these problems, as it aims to induce *high-quality* parallel sentences efficiently using automatic means. In our research, we begin by analysing the limitations of the existing DA methods and propose strategies to overcome those limitations, aimed at improving the NMT performance. To generalise our findings, we conduct this study on three language pairs: English-Sinhala, English-Tamil and Sinhala-Tamil. They belong to three distinct language families. Further, Sinhala and Tamil are known to be morphologically rich languages, making NMT further challenging.

First, we follow the word or phrase replacement-based augmentation strategy, where we induce *synthetic* parallel sentences by augmenting rare words and by using words from a bilingual dictionary. Our technique improves existing techniques by using both syntactic and semantic constraints to generate high-quality parallel sentences. This method improves translation quality for sequences containing out-of-vocabulary terms and yields better overall NMT scores than existing techniques. Secondly, we conduct an empirical study with three multilingual pre-trained language models (multiPLMs) and demonstrate that both the pre-training strategy and the nature of the pre-training data significantly affect the quality of mined parallel sentences. Thirdly, we enhance the cross-lingual sentence representations of existing encoder-based multiPLMs, in order to overcome their suboptimal performance in sentence retrieval tasks. We introduce *Linguistic Entity Masking* to enhance the cross-lingual representations of such multiPLMs and empirically prove that the improved representations lead to performance gains for cross-lingual tasks. Finally, we explore the Parallel Data Curation (PDC) task. In line with existing work, we identify that the scoring and ranking with different multiPLMs results in a disparity, which is caused by the multiPLMs' bias towards certain types of noisy parallel sentences. We show that multiPLM-based PDC, together with a heuristic combination, is capable of minimising this disparity while producing optimal NMT scores. Overall, we show that improving DA techniques leads to generating *high-quality* parallel data, which in turn leads to elevating the state-of-the-art benchmark NMT results further.



**Keywords**: Data Augmentation, Neural Machine Translation, Low Resource Languages, Bitext Mining, Parallel Data Curation

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ABBREVIATIONS

| Abbreviation | Description |
|---|---|
| AI | Artificial Intelligence |
| BPE | Byte-pair-Encoding |
| CNN | Convolutional Neural Network |
| DA | Data Augmentation |
| DAN | Deep Averaging Networks |
| DNN | Deep Neural Network |
| GRU | Gated Recurrent-Unit |
| HRL | High-Resource Languages |
| LLM | Large Language Models |
| LM | Language Model |
| LRL | Low Resource Languages |
| LRNMT | Low Resource Neural Machine Translation |
| LSTM | Long Short-Term Memory |
| MNMT | Multilingual Neural Machine Translation |
| MT | Machine Translation |
| multiPLM | Multilingual Pre-trained Language Model |
| NER | Named Entity Recognition |
| NET | Named Entity Translation |
| NLP | Natural Language Processing |
| NMT | Neural Machine Translation |
| OOV | Out-of-Vocabulary |
| PBSMT | Phrase-Based Statistical Machine Translation |
| POS | Part of Speech |
| RNN | Recurrent Neural Networks |

| Abbreviation | Description |
|---|---|
| RBMT | Rule-based Machine Translation |
| SMT | Statistical Machine Translation |
| UNMT | Unsupervised Neural Machine Translation |

# CHAPTER 1

# INTRODUCTION

## 1.1 Background and Motivation

In a world where global collaboration and information exchange are growing at an unprecedented pace, Machine Translation (MT), the automatic translation of text or speech from one language to another using computational methods, has emerged as a critical technological bridge to enable seamless communication across borders.

Neural Machine Translation (NMT) is the State-of-the-Art solution for the MT problem. NMT is an end-to-end trained deep neural network, based on an encoder–decoder architecture, that learns a sequence-to-sequence mapping from a source language to a target language. The transformer architecture (Vaswani et al., 2017) further elevated the benchmark scores significantly on NMT evaluation datasets. These systems, however, rely heavily on large-scale parallel datasets to train and produce state-of-the-art results. Parallel data refers to a collection of sentences in two languages where each sentence in one language is aligned with its corresponding translation in the other language(s). Such data is crucial for the NMT model to learn the sequence-to-sequence mappings between the two languages. Nevertheless, when parallel data is limited, particularly for Low-Resource Languages (LRLs), NMT models trained using the same architectures tend to produce sub-optimal results (Haddow et al., 2022; Ranathunga et al., 2021). LRLs are those that have limited or no linguistic resources and tool support (Ranathunga and de Silva, 2022).

Morphologically rich languages pose additional challenges to NMT. In such languages, words inflect based on factors like gender, number, and case, leading to a substantially larger vocabulary space. Many of these languages also fall into the category of LRLs (Haddow et al., 2022). Hence, for morphologically rich LRLs, the parallel data scarcity problem worsens the NMT performance. This degradation in performance can be attributed to several factors:

- **Vocabulary Limitations:** A limited parallel dataset covers only a subset of the vocabulary in both source and target languages. As a result, NMT systems struggle to translate sequences containing Out-of-Vocabulary (OOV) words, often leading to incorrect or inadequate translations.
- **Limited Contextual Diversity:** Even when words are present in the training corpus, their occurrence in limited and insufficiently diverse contexts limit the model from learning nuanced meanings.
- **Weak Sequence-to-Sequence Mappings:** The lack of sufficient parallel sentence pairs to learn the sequence-to-sequence mappings hinders the NMT model's

ability to establish strong syntactic and semantic alignments between source and target languages, further degrading the translation quality.

High-Resource Languages (HRLs) benefit from the availability of large-scale, high-quality parallel corpora, often compiled through human annotation efforts (Ziemski et al., 2016; Koehn, 2005) or using automatic techniques to induce parallel corpora from multilingual web sources (De Gibert et al., 2024; El-Kishky et al., 2020; Bañón et al., 2020). There's a long tail of 6500+ languages that fall into the category of LRLs or extremely low-resource languages (Ranathunga and de Silva, 2022), which are hindered by the parallel data scarcity problem (Haddow et al., 2022).

Data Augmentation (DA) aims at inducing parallel sentences following synthetically or by automatic means. Therefore it is a viable solution to address the parallel data scarcity problem for LRLs. We could primarily observe four DA techniques in the literature as word or phrase replacement-based augmentation (Duan et al., 2020; Peng et al., 2020; Fadaee et al., 2017), parallel sentence mining (bitext mining) from web crawled data (De Gibert et al., 2024; Costa-jussà et al., 2022; Schwenk et al., 2021b), parallel data curation (Sloto et al., 2023; Koehn et al., 2020) and back translation (Sennrich et al., 2016a; Fadaee and Monz, 2018). Bitext mining (Bañón et al., 2020) is a pipeline of subtaks where the objective is to produce parallel sentence-pairs from web-crawled comparable corpora. Two key subtasks which contributes to the quality of the resulting parallel sentences are document alignment and sentence alignment subtasks. State-of-the-art bitext mining techniques rely on embeddings generated by multiPLMs to identify aligned parallel documents or sentences. However, due to the under-representation of LRLs during the pre-training phase (Feng et al., 2022), the resulting multilingual embeddings for sentences, which are translations of each other, are not always positioned closely within the shared embedding space. This discrepancy leads to misalignments or the selection of sentence pairs that are only weakly equivalent in meaning. Additionally, the inherent noise in the available data, coupled with the sub-optimal performance of Natural Language Processing (NLP) tools, often results in poor-quality parallel datasets (Ranathunga et al., 2024a; Kreutzer et al., 2022a; Bane et al., 2022). Since NMT systems are sensitive to noise, training them with such noisy parallel data results in poor-quality translations.

## 1.2 Research Objectives

In this thesis, we aim to improve the existing DA techniques to produce *high-quality parallel sentences* to address both issues aforementioned in Section 1.1. We direct this research under four Research Objectives, detailed in this section. Figure 1.1, is a summary of the work we have done. It outlines the gap in literature on which we have based each Research Objective, the contributions made after addressing each research objective and the publication details.

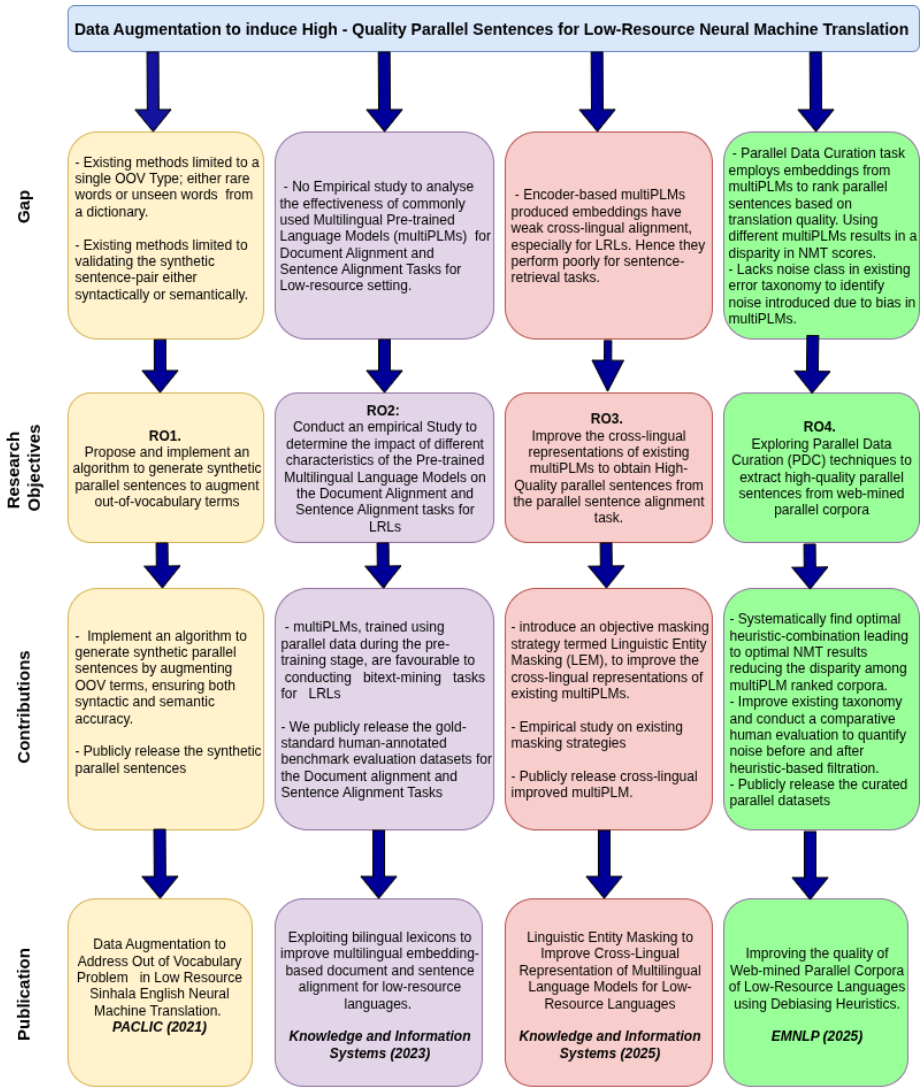


Figure 1.1: Research objectives vs contributions and publication mapping

**RO1. Propose and implement an algorithm to generate synthetic parallel sentences to augment out-of-vocabulary terms.**

We first explore the word or phrase replacement-based DA approach, which is aimed at generating synthetic parallel sentence pairs (Fadaee et al., 2017) by modifying the existing parallel sentences semantically while preserving the syntactic and semantic coherence in the resulting sentences. We decide to augment Out-of-Vocabulary (OOV) (Sennrich et al., 2016b) terms in our augmentation. OOV are two fold: They can be rare words (Fadaee et al., 2017), which are words that occur in low frequency in the training corpus or unseen words (Peng et al., 2020), which are words that do not exist in the training corpus at all. Since such words are not encountered by the NMT model during training, the presence of OOV words in the input sentence often results in poor translation quality. Secondly, we observe that existing techniques are limited to validating the final synthetic sentences for either syntactic correctness (Peng et al., 2020) or semantic correctness (Tennage et al., 2018b). Therefore, in our work,

we augment both types of OOV terms and take measures to validate the plausibility of the resulting synthetic sentences using a series of syntactic and semantic constraints. Our hypothesis is that the generated synthetic parallel sentences lead to improving the translations of the sequences containing OOV terms, as well as to improving the overall NMT scores between the language pairs.

**RO2: Conduct an empirical Study to determine the impact of different characteristics of the Pre-trained Multilingual Language Models on the Document Alignment and Sentence Alignment tasks for LRLs.**

Bitext mining (Costa-jussà et al., 2022; Gala et al., 2023; Bañón et al., 2020) alleviates the parallel data scarcity problem even for LRLs by extracting *parallel sentences* from *comparable corpora*. The bitext mining pipeline consists of several subtasks, among which web crawling, document alignment, and sentence alignment are crucial for ensuring the quality of the extracted parallel sentences. State-of-the-art approaches for these tasks (El-Kishky et al., 2020; Schwenk et al., 2021b,a) include obtaining sentence representation for the documents or sentences using encoder-based multiPLMs, and determining the alignment, based on semantic similarity scoring. However, under representation of LRLs in the training data (Conneau et al., 2020a) during the pre-trained stage and the pre-trained objective (Conneau and Lample, 2019) impact the reliability of the resulting vector representations for *cross-lingual* tasks. In this study, we conduct an empirical evaluation of the effectiveness of commonly used, multiPLMs for the document alignment and sentence alignment tasks in a LRL setting. Thereby, we gain insights into the characteristics inherent to the multiPLM that lead to optimal results for the considered tasks.

**RO3. Improve the cross-lingual representations of existing multiPLMs to obtain High-Quality parallel sentences from the parallel sentence alignment task.**

MultiPLMs (Conneau et al., 2020a; Devlin et al., 2019a), trained using the Masked Language Modeling (MLM) objective, are widely used for *cross-lingual* tasks. However, their performance remains sub-optimal for sentence-retrieval tasks (Hu et al., 2021a; Feng et al., 2022) due to the lack of an explicit cross-lingual objective. Conneau and Lample (2019) have demonstrated that a MLM pre-trained model can be continually trained in a Translation Language Modelling (TLM) step, to enhance the alignment of the cross-lingual representations, produced in the MLM step. The TLM step utilize parallel data, and is driven by the intuition that the prediction of the masked tokens would be attributed not only by its context, but also by its translation counterpart. However, both MLM and TLM select tokens for masking randomly, disregarding the linguistic properties of individual tokens. We argue that such randomness does not lead to producing optimal *sentence representations*, to be favourable for downstream NLP tasks. Hence, we hypothesise that masking linguistically informed tokens will optimise the cross-lingual representation improvement further.

**RO4. Exploring Parallel Data Curation (PDC) techniques to extract high-quality parallel sentences from web-mined parallel corpora.**
Parallel Data Curation (PDC) (Sloto et al., 2023; Koehn et al., 2020, 2019, 2018), techniques aim to filter out noisy parallel sentences from web-mined corpora to produce *high-quality* data for training NMT systems. Initiated by the work of Chaudhary et al. (2019), recent PDC techniques (Steingrímsson, 2023; Gala et al., 2023) follow a scoring and ranking mechanism using embeddings obtained from a multiPLM. Then, a subset of the top-ranked sentence pairs is selected to train the NMT model. However, the quality of these top-ranked pairs depends on the selected multiPLM (Ranathunga et al., 2024a; Moon et al., 2023) and leads to a performance disparity in the trained NMT models. The disparity is caused due to the inherent bias in the multiPLMs, prioritizing certain characteristics of parallel sentences, which are actually noise from a NMT model point of view. We hypothesize that this disparity can be mitigated with heuristics and aim to conduct an ablation study to identify the heuristic combination leading to optimizing the NMT scores as well as minimizing the disparity.

## 1.3 Contributions

This thesis makes several key contributions, including the release of artifacts such as data, improved models, benchmark evaluation sets and the dissemination of findings from the scientific studies conducted as part of this research. Most of these have been published in peer-reviewed conferences and journals. They are discussed in this section. First, we describe the key scientific contributions of this research:

- We propose an algorithm to generate synthetic parallel sentences by augmenting Out-of-Vocabulary (OOV) terms, which exist as rare words and dictionary terms. To ensure that high-quality synthetic parallel sentences are produced, we employ both syntactic and semantic features to validate the final synthetic sentences for correctness. (Fernando and Ranathunga, 2021)

- We conduct an empirical study to evaluate the effectiveness of multiPLMs for the document alignment and sentence alignment tasks, in the bitext mining pipeline. We show that multiPLMs, trained using parallel data during the pre-training stage, are favourable to conducting bitext-mining tasks for LRLs. (Fernando et al., 2023)

- We introduce an objective masking strategy termed Linguistic Entity Masking (LEM), to improve the cross-lingual representations of existing multiPLMs. We empirically show that existing masking strategies are less effective on the sentence alignment task and that LEM is more favourable to improving the performance of the sentence alignment task for morphologically rich LRL language-

pairs English-Sinhala, English-Tamil and Sinhala-Tamil. (Fernando and Ranathunga, 2025)

- In line with the existing work, we show that using different multiPLMs for ranking the parallel sentence pairs, and by training the NMT systems using the top-ranked parallel sentences, results in a disparity. We systematically evaluate the impact of heuristics in the PDC step and empirically show that it leads to reducing the disparity among the NMT systems drastically. Based on this study, we propose a heuristic-based PDC to be applied before ranking parallel sentences for training NMT systems, especially in the case of web-mined corpora. (Fernando et al., 2025)

- We introduce a noise category for the existing noise taxonomy by (Ranathunga et al., 2024a) and improve the definitions of two existing noise categories. (Fernando et al., 2025)

- We conduct a human evaluation to quantify the noise in the top-ranked parallel data and outline the bias pertaining to each multiPLM. Subsequently, after the heuristic-based PDC step, we show that the quality of the top-ranked parallel corpora becomes comparable. (Fernando et al., 2025)

Next, we describe the data, models and resources released for the research community.

- We release the synthetic parallel sentences, augmenting both rare words and words from a bilingual dictionary publicly.[1] (Fernando and Ranathunga, 2021).

- We publicly release the gold-standard human-annotated benchmark evaluation datasets for the document alignment[2] and sentence alignment[3] tasks for three low-resource language pairs: English-Sinhala, English-Tamil and Sinhala-Tamil. This is the only benchmark evaluation dataset available for these language pairs. (Fernando et al., 2023)

- We have improved the cross-lingual representations of XLM-R, an encoder-based multiPLM using the LEM strategy for the three language-pairs under our study: English-Sinhala, English-Tamil and Sinhala-Tamil. We release these models [4] publicly to be used not only for sentence-retrieval tasks, but also for the benefit of NLP tasks which require cross-lingual representations. (Fernando and Ranathunga, 2025)

[1] https://github.com/aloka-fernando/Inducing-Synthetic-Parallel-Sentences

[2] https://huggingface.co/datasets/NLPC-UOM/document_alignment_dataset-Sinhala-Tamil-English

[3] https://huggingface.co/datasets/NLPC-UOM/sentence_alignment_dataset-Sinhala-Tamil-English

[4] https://github.com/aloka-fernando/Linguistic-Entity-Masking-LEM

- We release the curated CCMatrix and CCAligned datasets for the three language pairs.[5] (Fernando et al., 2025)

## 1.4 Structure of the Thesis

In Chapter 2, we review prior work on advancements in MT architectures, summarize key techniques for training NMT systems, and discuss existing data augmentation methods designed to enhance low-resource NMT. We also outline commonly used benchmark datasets and evaluation metrics that are relevant to the work done in this thesis.

In Chapter 3, we present our approach to synthetic parallel-sentence generation, which is motivated by the word or phrase-based augmentation technique. In this, we augment rare words and unseen words from a bilingual dictionary and prove that the synthetic parallel sentences lead to improving the overall NMT score and the translation quality for the sequences containing these OOV terms.

In Chapter 4, we consider the bitext mining approach and conduct an empirical study to analyse the effectiveness of the commonly available multiPLMs in the document alignment and sentence alignment sub-tasks. Then we present our empirical findings.

In Chapter 5, we present the Linguistic Entity Masking (LEM) strategy proposed to improve the cross-lingual embeddings of existing multiPLMs to optimize sentence retrieval tasks, mainly targeting the sentence alignment task.

In Chapter 6, we summarize our work on Parallel Data Curation (PDC), focusing on the use of multiPLMs for ranking parallel sentences to filter noisy sentence pairs in web-mined corpora. We highlight that the variation in curation quality is due to the biases inherent to the multiPLMs and empirically show that applying heuristics can mitigate this disparity. We further present the human evaluation conducted along with the results to show that, after applying heuristic-based PDC, qualitatively the curated corpora are comparable.

Finally, conclusions are given in Chapter 7 where we summarize our findings and point out possible future directions for our work.

[5] https://github.com/aloka-fernando/Heuristic-based-Parallel-Data-Curation

## 1.5 Publications

The contributions and findings discussed above have been published in peer-reviewed conferences and journals:

- **Fernando, A.** Ranathunga, S., & de Silva, N. (2025). Improving the quality of Web-mined Parallel Corpora of Low-Resource Languages using Debiasing Heuristics. arXiv preprint arXiv:2502.19074. (Accepted at The 2025 Conference on Empirical Methods in Natural Language Processing)
  **Core Rank: A*/ h-Index: 193**

- **Fernando, A.** & Ranathunga, S. (2025). Linguistic entity masking to improve cross-lingual representation of multilingual language models for low-resource languages. Knowl Inf Syst (2025).
  **Qartile: Q2; h-Index: 100**

- **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. Knowledge and Information Systems, 65(2), 571-612.
  **Qartile: Q2; h-Index: 100**

- **Fernando, A.**, & Ranathunga, S. (2021). Title: Data Augmentation to Address Out of Vocabulary Problem in Low Resource Sinhala English Neural Machine Translation. In Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation (pp. 61-70).
  **h5-Index: 13**

## 1.6 Other Publications

I have contributed to the following publications during my PhD study period. They do not come under the scope of study in this research (i.e. they are not outputs of my PhD research). My contributions to them have been by conducting data analysis (Ranathunga et al., 2023), contributing to the data preparation phase (Velayuthan et al., 2024), corpus development (Ranathunga et al., 2024a; Fernando et al., 2020) and linguistic resource compilation (Fernando and Dias, 2021).

- Velayuthan, M., Jayakody, D., De Silva, N., **Fernando, A.**, & Ranathunga, S. (2024, November). Back to the Stats: Rescuing Low-Resource Neural Machine Translation with Statistical Methods. In Proceedings of the Ninth Conference on Machine Translation (pp. 901-907).
  **h-Index: 45**

- Ranathunga, S., De Silva, N., Jayakody, D., & **Fernando, A.** (2024, August). Shoulders of Giants: A Look at the Degree and Utility of Openness in NLP Research. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 519-529).
  **h-Index: 215**

- Ranathunga, S., De Silva, N., Menan, V., **Fernando, A.**, & Rathnayake, C. (2024, March). Quality Does Matter: A Detailed Look at the Quality and Utility of Web-Mined Parallel Corpora. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 860-880).
  **Best Paper Award in Low-Resource Category; h-Index: 56**

- **Fernando, A.**, & Dias, G. (2021, December). Building a Linguistic Resource: A Word Frequency List for Sinhala. In Proceedings of the 18th International Conference on Natural Language Processing (ICON) (pp. 606-610).

- **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.

## 1.7 Definitions

In this section, we introduce and define the terminology that will be used extensively throughout this thesis.

MT systems accept a sentence in the source language and output the translation of it in the target language. This output translation is expected to be of *high-quality*, which is defined by *adequacy* and *fluency*. Adequacy refers to how well the translation preserves the exact meaning of the source sentence. Fluency measures the grammatical correctness, clarity, readability, and naturalness of the translation in the sentence. If the translation output lacks any of these properties, the MT system is said to be *sub-optimal*.

MT systems are trained using parallel sentence pairs. A *parallel sentence pair* (Tiedemann, 2012; Heffernan et al., 2022) consists of a source language sentence and a target language sentence that are semantically equivalent and aligned at the sentence level. A *high-quality parallel sentence pair* (Ranathunga et al., 2024a) is when the source and target sentences are perfect translations of each other and excel in both adequacy and fluency. A *noisy parallel sentence pair* (Khayrallah and Koehn, 2018) exhibits errors such as mistranslations, misalignment, incomplete translations, or unnatural phrasing. Such pairs can degrade the quality of MT models. *Parallel data* refers to a collection of texts in two or more languages in which each sentence in one language is aligned with its corresponding translation counterpart in the other language. *Alignment* refers to

the mapping of linguistic units, such as words or phrases, between the source language sentence and the target language sentence.

The techniques employed in this thesis primarily rely on the notion of *sentence representations* or *vector representations* (Reimers et al., 2019), where each sentence is encoded into a fixed-dimensional vector in the embedding space. These representations are typically obtained using *multiPLMs* (Wang et al., 2025), which are neural models trained on large-scale multilingual corpora to capture both syntactic and semantic information across multiple languages. A key property leveraged in this work is the ability of such models to generate *cross-lingual embeddings* (Kreutzer et al., 2022a), which are representations in a shared embedding space where semantically similar sentences from different languages are closely aligned. These embeddings form the basis for *sentence retrieval tasks* (Leveling et al., 2012), which involve identifying the semantically equivalent sentence(s) from candidate sentences in the target language for a given source sentence or vice versa.

# CHAPTER 2

# BACKGROUND

In this chapter, we present key developments in Machine Translation (MT) research. We begin by discussing the architectural evolution of MT systems, tracing the progression from rule-based systems to the modern transformer-based architectures. This is followed by an overview of the main methodological approaches that have shaped MT over time. We then shift our focus to Low-Resource Neural Machine Translation (LRNMT), highlighting techniques that have proven effective in such settings. In particular, we explore data augmentation as a promising research direction for enhancing translation quality in low-resource scenarios. Finally, we provide an overview of selected low-resource languages, examine the availability of relevant datasets, and conclude with a brief discussion on the evaluation metrics in NMT, laying the groundwork for the research outlined in the subsequent chapters.

## 2.1 Machine Translation

Machine Translation research extends to handling different input modalities such as text, speech or sign language. This thesis addresses MT as a sequence-to-sequence problem involving sentence-level textual translations. The concept of MT was formally introduced by Weaver (1955), who envisioned the use of modern computers to automate the translation of human languages. Since then, MT has been considered one of the most challenging problems in the fields of Natural Language Processing (NLP) and Artificial Intelligence (AI).

## 2.2 Machine Translation Techniques

Over the years, various techniques have been developed to address the MT problem. These approaches can be broadly categorized as illustrated in Figure 2.1.

### 2.2.1 Rule-based Machine Translation

Rule-based Machine Translation (RBMT) was the earliest approach for automating translation. It relied on linguistic rules, dictionaries, and grammars handcrafted by experts (Yazar et al., 2023). RBMT systems were described as a three-phase architecture, dividing translation into analysis, transfer, and generation (Nagao and Tsujii, 1986; Allen B. Tucker and Nirenburg, 1984; Nagao et al., 1980). Subsequent work introduced modularity to separate linguistic knowledge among components (Carlson and Vilkuna, 1990; Isabelle and Macklovitch, 1986), employed pattern-based, paired grammar rules (Novák et al., 2008; Prószéky, 2005) and acquired rules by semi-automatic

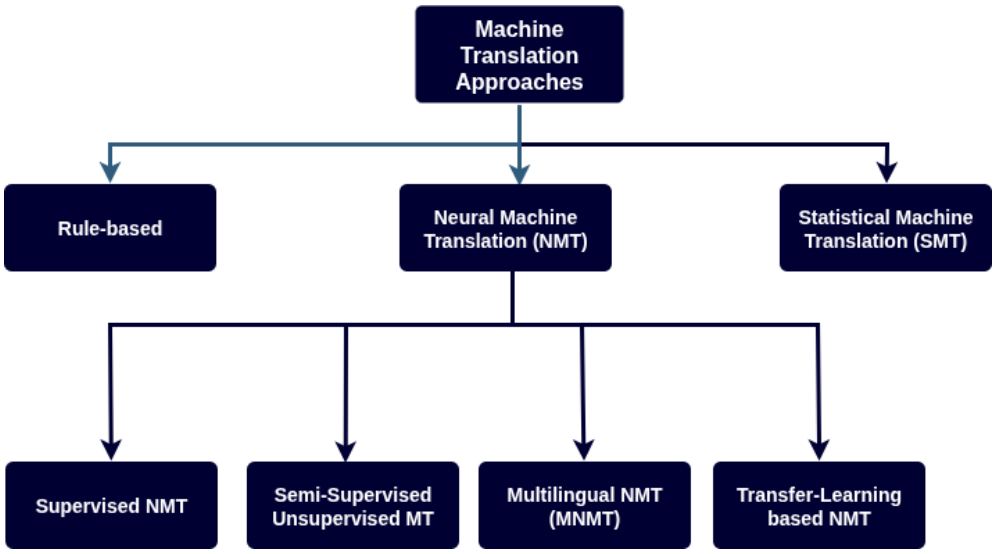

Figure 2.1: Classification of MT Techniques

means (Winiwarter, 2007). A major setback of RBMT systems (Stojanovski, 2021) was their reliance on extensive, manually crafted rules, which made them inherently complex and time-consuming to develop.

### 2.2.2 Statistical Machine Translation

Statistical Machine Translation (SMT) emerged as a probabilistic framework based on a noisy channel model (Brown et al., 1993). The word-based SMT was based on Bayes' theorem, where the translation probability for translating source sentence $f$ into target sentence $e$ was represented as in Equation 2.1. Here, $p(f|e)$ represented the translation model probability and $p(e)$ represented the Language Model (LM) probability.

$$\hat{e} = \arg\max_{e} P(e \mid f) = \arg\max_{e} P(f \mid e) \times P(e) \tag{2.1}$$

The objective of the popular Phrase-Based Statistical Translation Model (PBSMT) (Koehn et al., 2003) maximized the probability considering several compositional models. Here the probabilities were derived from $P_\phi(f \mid e)$ phrase translation model, $P_{\text{LM}}(e)$ Language Model and a $P_D(e, f)$ distortion model (Och and Ney, 2003) respectively. $\omega^{\text{length}(e)}$ was a length penalty.

$$\hat{e} = \arg\max_{e} P_\phi(f \mid e) \times P_{\text{LM}}(e) \times P_D(e, f) \times \omega^{\text{length}(e)} \tag{2.2}$$

In the PBSMT, a phrase table containing the translation probabilities was constructed from a parallel corpus, considering a word or phrase alignment model. The reordering model (also called the distortion model) indicated where to place translated phrases in the target sentence, especially when the word order differs between source and target languages. Additionally, an n-gram language model determined the fluency of

the target translation. The length penalty was incorporated into the SMT decoder's scoring function to compensate for its tendency to favour shorter translations. The SMT dominated the MT research for nearly three decades with the availability of the open source toolkit such as Moses (Koehn et al., 2007).

### 2.2.3 Neural Machine Translation (NMT)

In this section, we describe the neural network architectures which are prominent in training translation models.

#### 2.2.3.1 RNN-based NMT Architecture

Neural Machine Translation (NMT) represented a paradigm shift. It refers to training a Neural Network to produce a translation in the target language for the given source language sentence. The early works of NMT by Kalchbrenner and Blunsom (2013), modelled the NMT using a Convolution Neural Network (CNN) as an encoder and a Recurrent Neural Network (RNN) as a decoder. Subsequently, Cho et al. (2014) and Sutskever et al. (2014) used RNNs for both encoder and decoder, where the encoder represented the variable-size sequence into a fixed-size context vector and the decoder converted this into a sequence in the target language. The two networks were jointly trained to maximize the conditional probability of the target sequence given a source sequence. However, the former used RNN-based sequence-to-sequence model to learn the phrase translation probabilities to improve the PBSMT using Gated Recurrent Units (GRUs), while the latter used Long-Short Term Memory (LSTM) units.

RNN (Werbos, 1990) is a feedforward Deep Neural Network (DNN) for handling variable-size sequences. Given a sequence of inputs $\mathbf{X} = (x_1, x_2 ... x_T)$, a standard RNN computed a sequence of outputs $\mathbf{Y} = (y_1 y_2 ... y_T)$ by iterating the following equations (Cho et al., 2014):

$$h_t = \phi\left(W^{hx} x_t + W^{hh} h_{t-1}\right) \tag{2.3}$$

$$y_t = W^{yh} h_t \tag{2.4}$$

Equation 2.3 shows the hidden state $h_t$ at each timestep $t$ updated as a function of the previous hidden state $h_{t-1}$ and the current input $x_t$ by the RNN. Here, $W^{hx}$ and $W^{hh}$ were weight matrices. $\phi$ can be any non-linear activation function.

Using LSTM (Sutskever et al., 2014) as a RNN unit replaced the hidden state at each timestep with a long-term hidden state, which better represented the input sequence, capturing the long-term dependencies. This enabled the LSTM units to optimize the conditional probability $p(y_1 y_2 ... y_{T'} | x_1 x_2 ... x_T)$ of the output sequence

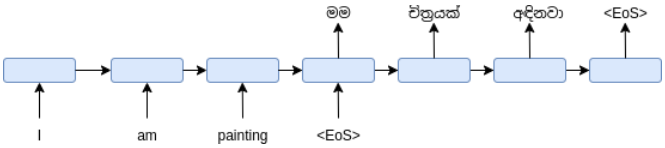


Figure 2.2: Translation example with RNN. Adapted from (Sutskever et al., 2014)

defined in Equation 2.5. Here, $v$ was the fixed-size content vector, which was the final long-term state from the encoder.

$$p(y_1, \ldots, y_{T'} \mid x_1, \ldots, x_T) = \prod_{t=1}^{T'} p(y_t \mid v, y_1, \ldots, y_{t-1}) \tag{2.5}$$

The fixed-size context vector was a bottleneck in the RNN-based encoder-decoder architecture, as it was expected to capture all relevant information from the input, regardless of its length or complexity. This limitation was overcome by the attention mechanism that was proposed by Bahdanau et al. (2015). In this architecture, two differences were prominent. Firstly, they used a bidirectional RNN as the encoder.

Secondly, during decoding, the model computed a set of attention weights $\alpha_{t,i}$ over all source positions for each target token $y_t$. These weights indicated the source token $x_i$ when generating $y_t$. The attention weights were computed using an alignment function as in Equation 2.6.

$$p(y_i \mid y_1, \ldots, y_{i-1}, x) = g(y_{i-1}, s_i, c_i) \tag{2.6}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \quad \text{where} \quad e_{ij} = a(s_{i-1}, h_j) \tag{2.7}$$

Here, $s_{t-1}$ was the decoder hidden state from the previous time step, and $h_i$ was the encoder hidden state at source position $i$. The context vector $\mathbf{c}_t$ for time step $t$ was then computed as a weighted sum of the encoder hidden states:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \tag{2.8}$$

This context vector $\mathbf{c}_t$ was used alongside the decoder state $s_{t-1}$ to generate the target word $y_t$. In essence, the attention mechanism allowed the decoder to emulate searching through the source sentence to find the most relevant information for generating each target token.

Luong et al. (2015) further refined attention by comparing global and local alignment strategies. These developments significantly improved translation quality in supervised settings by making training more effective and outputs more contextually accurate.

#### 2.2.3.2 Transformer-based NMT Architecture

The Transformer model by (Vaswani et al., 2017) is the State-of-the-Art architecture for NMT systems. Moving away from recurrence, it introduced a fully attention-based architecture with self-attention, multi-head mechanisms, and positional encoding. As shown in Figure 2.3, the encoder includes identical layers (N) and each layer is composed of two sub-layers: the self-attention sub-layer followed by the feed-forward sub-layer.

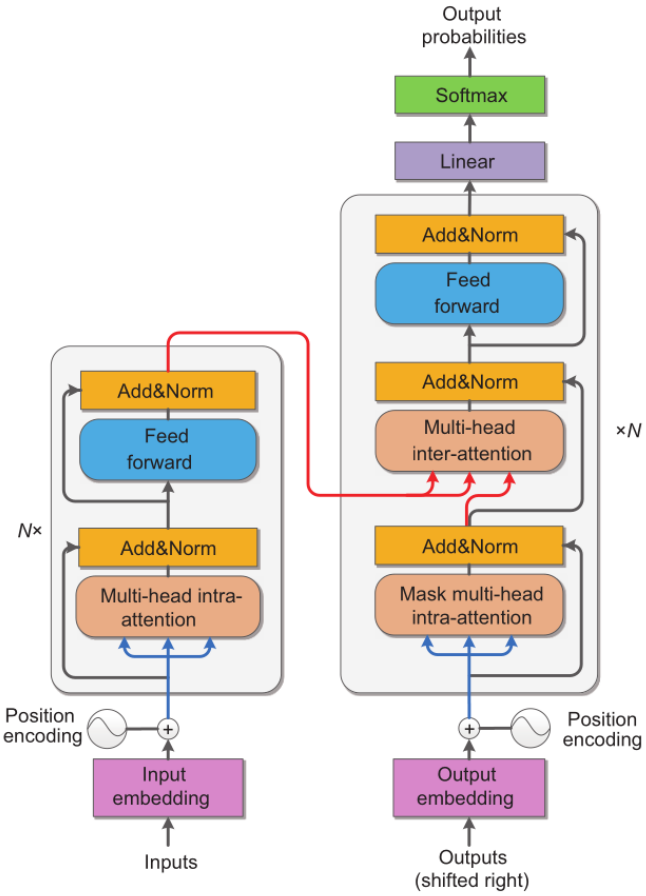


Figure 2.3: Transformer Architecture. Adapted from Zhang and Zong (2020)

The self-attention sub-layer computes a token's contextualised representation by attending to all other tokens within the same layer. It does so by calculating correlation scores between the source side token embeddings and their surrounding embedding, and then producing a weighted sum of all embedding representations, including the embedding itself. The output of the final *(Nth)* encoder layer serves as the source-side semantic representation, denoted by $h$. Decoder also contain identical layers. Each layer contains three sub-components: (1) a masked self-attention sub-layer that captures the partial prediction history, (2) an encoder-decoder attention sub-layer that dynamically attends to the encoder outputs based on the current decoding context, and

(3) a feed-forward sub-layer.

Despite their roles, all three attention mechanisms can be unified under a common mathematical formulation in Equation 2.9.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2.9}$$

Here, Q, K and V stand for a query, the key list and the value list, respectively. $d_k$ is the dimension of the key. For the encoder self-attention, the queries, keys and values are from the same layer. When calculating the output of the first layer in the encoder at the *jth* position, let $x_j$ be the sum vector of the input token embedding and the positional embedding. The query is vector $x_j$. The keys and values are the same, and both are the embedding matrix $X = [x_0...x_j]$. Then, multi-head attention is proposed to calculate attentions in different subspaces.

$$\begin{aligned}\text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \\ \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)\end{aligned} \tag{2.10}$$

Here, $W_i^Q$, $W_i^K$, $W_i^V$, and $W^O$ denote the projection parameter matrices used to transform the input queries, keys, and values, as well as to project the concatenated attention heads to the output space, respectively. Using Equation (2.10), followed by a residual connection, layer normalization, and a position-wise feed-forward network, we obtain the output representation of the second layer. By repeating this process over $N$ layers, we arrive at the final encoder output representations, denoted as the input context. Here, $\mathbf{h}_J$ represents the contextualized embedding at position $J$.

$$C = [\mathbf{h}_0, \mathbf{h}_1, \ldots, \mathbf{h}_J] \tag{2.11}$$

The masked self-attention mechanism in the decoder is similar to the encoder's self-attention, with the key distinction that, at position $i$, the query vector is restricted to attend only to positions $\leq i$. This constraint ensures that the model does not access future positions during auto-regressive left-to-right decoding. The masked attention at position $i$ is formulated as:

$$\mathbf{z}_i = \text{Attention}(\mathbf{q}_i, K_{\leq i}, V_{\leq i}) = \text{softmax}\left(\frac{\mathbf{q}_i K_{\leq i}^\top}{\sqrt{d_k}}\right)V_{\leq i} \tag{2.12}$$

The encoder-decoder attention mechanism computes the dynamic source-side context required for predicting the current target token. In this case, the query is the output from the masked self-attention sub-layer, $\mathbf{z}_i$, while both the keys and values are the encoder outputs $C = [\mathbf{h}_0, \mathbf{h}_1, \ldots, \mathbf{h}_J]$.

Following this attention mechanism, the residual connection, layer normalisation, and feed-forward sub-layer are applied sequentially to produce the output of the current

decoder layer. After $N$ such stacked layers, we obtain the final decoder hidden state $\mathbf{z}_i$. A softmax layer is then applied to generate the probability distribution over the target vocabulary and predict the output token $y_i$, as illustrated in the upper-right portion of Figure 2.3.

## 2.3 Prominent Techniques in NMT

As shown in Figure 2.1, in this section, we outline the prominent techniques which had evolved related to NMT.

**Supervised NMT** is when the NMT model is trained, using the sequence-to-sequence training examples given in terms of parallel sentences. Advancement of supervised NMT was mainly due to the availability of large-scale parallel corpora (Tan et al., 2019; Tiedemann, 2012). The supervised sentence-level NMT, was extended to subsequently to translation at the paragraph-level (Araabi and Monz, 2020) and document-level (Zhang et al., 2018; Lopes et al., 2020). To this date, supervised NMT systems trained on large-scale, high-quality parallel sentences yield state-of-the-art performance for high-resource (Takase and Kiyono, 2023) language pairs.

Unsupervised Neural Machine Translation (UNMT) has emerged as a transformative approach in MT by enabling systems to be trained solely on monolingual corpora, without relying on parallel data. In UNMT, word embeddings are trained using the monolingual data for the respective languages, then embeddings are aligned using auto-encoders between the two languages. (Lample and Conneau, 2018). It is favourable for LRLs when the parallel data is scarce or does not exist. Early works by Lample and Conneau (2018); Artetxe et al. (2018a) demonstrated that shared encoder representations, denoising objectives, and iterative back-translation techniques could produce UNMT systems even between distinct language pairs. Conneau and Lample (2019) proposed XLM, introducing an unsupervised cross-lingual language modelling objective to further enhance performance in both UNMT and cross-lingual understanding tasks. Subsequent advancements focused on improving alignment quality (Sun et al., 2021), adapting a contrastive learning objective. More recent research has explored aligning linguistic structures using self-supervised signals (Yang et al., 2023), and exploiting pre-trained multilingual encoders for domain-robust translation (Garcia et al., 2023). Collectively, these studies underscore the viability of UNMT and also cross-lingual generalization in the absence of parallel corpora.

**Multilingual Neural Machine Translation (MNMT)** research takes the direction of training a single, MNMT (Johnson et al., 2017; Ha et al., 2016) model to produce translations among many-to-many languages. Utilising parallel data from both HRLs and LRLs, MNMT model is aimed at sharing linguistic knowledge by means of parameter sharing, improving the translations for LRLs. While there can be many-to-one and one-to-many architectural variants in MNMT, key strategies employed

in MNMT include multi-task training (Xue et al., 2021a), multistage fine-tuning using auxiliary languages (Dabre et al., 2019) and shared encoder-decoder frameworks with multilingual vocabularies (Tars et al., 2022). NLLB-200 (Costa-jussà et al., 2022), M2M-100 (Gowda et al., 2021), mT5 (Xue et al., 2021a) and mBART (Tang et al., 2021) are MNMT models contributing to the success of this research domain. These methods have been shown to significantly improve LRNMT (San et al., 2024; Mager et al., 2023). However, challenges such as catastrophic forgetting and language imbalance are issues in MNMT (Roy et al., 2024).

**Transfer learning-based NMT** has elevated NMT benchmark scores, enabling models to generalise across tasks and languages by reusing knowledge from Pre-trained Language Models (PLMs). Early work in sequence-to-sequence transfer learning demonstrated that initialising low-resource NMT models with parameters from high-resource language pairs significantly improved translation quality (Zoph et al., 2016).

Foundation models such as mBART (Liu et al., 2020) and mT5 (Xue et al., 2021b) have been mostly considered as the parent model in Transfer Learning . These foundation models which leverage denoising and text-to-text objectives across many languages. More recently, generative Large Language Models (LLMs) or decoder-based pre-trained and fine-tuned models have been explored for zero-shot and few-shot translation capabilities, with models such as mGPT (Shliazhko et al., 2022) and BLOOM (Scao et al., 2022) showing promising results. These advancements highlight the role of Transfer Learning in closing performance gaps for LRLs and enabling MT systems.

In summary, these advances illustrate a clear trajectory in NMT research: moving from data-dependence to data-efficient learning, from dedicated bilingual systems to scalable multilingual frameworks, and from task-specific models to general-purpose PLMs capable of cross-lingual transfer.

## 2.4 Low-Resource Neural Machine Translation

LRNMT research objective is to improve existing NMT systems to provide reliable translation in a setting where parallel data is scarce or unavailable for the respective language pairs (Tan et al., 2019). Still, a long tail of 6500+ languages (Ranathunga and de Silva, 2022) fall into the category of LRL or extremely LRL languages and hence LRNMT research (Sánchez-Martínez et al., 2024; Weller-Di Marco and Fraser, 2022) is still relevant to this date. From the aforementioned techniques, UNMT, MNMT and most dominantly, transfer-learning based approaches benefit the LRNMT (Costa-jussà et al., 2022; Heffernan et al., 2022).

Nevertheless, there are challenges associated with LRNMT which still hinder the progression in this direction. Lack of parallel datasets and the attempts to augment parallel data results in low-quality parallel data (Latief et al., 2024; Ranathunga et al., 2023), issues arising due to relying on linguistically independent BPE tokenisation (Shi

et al., 2022; Liu et al., 2019), understudy of active learning or data selection impact (Ranathunga et al., 2023) and cross-domain or cross-language generalization (Shi et al., 2022; Dabre et al., 2020) are few from the list of open issues in LRNMT.

In this thesis, we focus on addressing the research gaps in the existing literature which are associated with data augmentation techniques to improve the benchmark scores of LRNMT.

## 2.5 Data Augmentation

The translation quality of NMT systems depends on the quality and quantity of the parallel data on which they are trained on. However, for LRLs, the training dataset size limitation and the unbalanced nature of the dataset would lead to model overfitting, leading the model to produce poor translations (Wang et al., 2024). In such instances, DA addresses this issue by generating more parallel sentence-pairs by automatic means (Zhou et al., 2024; Chen et al., 2023). In the context of NMT systems, these DA techniques can be categorised into four main areas (Haddow et al., 2022; Ranathunga et al., 2021) as shown in Figure 2.4.

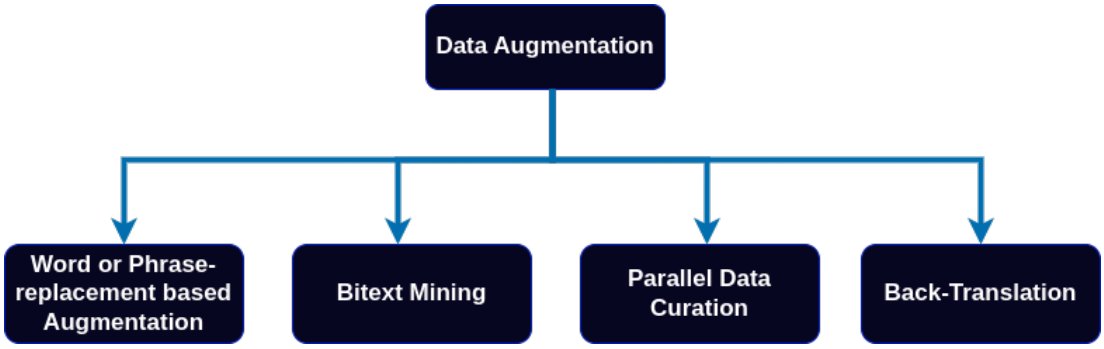


Figure 2.4: Classification of Data Augmentation Techniques in NMT

### 2.5.1 Word or Phrase Replacement-based augmentation

Word or phrase replacement-based DA (Fadaee et al., 2017; Tennage et al., 2017) is an effective strategy for improving NMT in low-resource settings by synthetically inducing parallel sentences. These methods generate semantically diverse synthetic parallel sentences by replacing low-frequency words in the training corpus or words totally absent from the training corpus. These techniques rely on linguistic tools (Nagy et al., 2023; Liu et al., 2023; Tennage et al., 2017) for syntactic information or word-embeddings (Peng et al., 2020) for semantic cues. Fadaee et al. (2017) and Peng et al. (2020) creates a LR scenario by using English-German and English-Frenth language pairs, while Tennage et al. (2018b) conducts their work considering Sinhala-Tamil

language pair. This line of research continues to be highly effective in LRL settings to this day (Liu et al., 2024; Ramesh et al., 2021b; Sen et al., 2021).

### 2.5.2 Bitext Mining

Bitext mining (Bañón et al., 2020) refers to the sequence of subtasks, web-crawling, document alignment and sentence alignment, which is aimed at inducing parallel sentence pairs from the multilingual or bilingual web-crawled comparable corpora. Recent approaches rely on cross-lingual or multilingual sentence embeddings to assess the semantic equivalence between the aligned source and target document-pairs or sentence-pairs. Bitext mining has been approached as Local mining (Bañón et al., 2020; Açarçiçek et al., 2020), where the document alignment is conducted prior to the sentence alignment task or as Global Mining (Costa-jussà et al., 2022; Schwenk et al., 2021b), where the document alignment is omitted, focusing on maximising the recall. The former is observed to produce well-aligned parallel sentences compared to the latter. The work under bitext mining covers a wide spectrum of languages, from HRLs to LRLs.

### 2.5.3 Parallel Data Curation

PDC techniques (Sloto et al., 2023; Koehn et al., 2020, 2019) are aimed at extracting *high-quality* parallel sentences from existing noisy parallel corpora (Ranathunga et al., 2024a; Bane et al., 2022). These approaches demonstrate that filtering is not only essential for mitigating the risks of noisy data in LRL setups but also for maximising the utility of augmented corpora (Steingrímsson, 2023; Minh-Cong et al., 2023b). Using automatic means to induce parallel data or when obtaining synthetic parallel data, the PDC step is essential. It was proven the NMT systems trained with the curated parallel dataset often outperforms the NMT systems trained with the full parallel dataset. The PDC work has been conducted mainly on LRLs, such as Estonian-Lithuanian (Sloto et al., 2023), Khmer-English (Koehn et al., 2020), Pashto-English (Koehn et al., 2020), Nepali-English (Koehn et al., 2019) and Sinhala-English (Koehn et al., 2019).

### 2.5.4 Back-Translation

Back-translation (Sennrich et al., 2016a) leverages monolingual target data and a NMT trained in the target-to-source direction to induce source-side sentences to obtain a parallel corpus for that language pair. Subsequent work involved selecting quality parallel sentences by means of iterative-back-translation (Epaliyana et al., 2021), coupling back-translation with self-training (Abdulmumin et al., 2020) and leveraging MNMT systems to generate back-translated data (Lu et al., 2024). Still in the context

on LRNMT, since the NMT systems are sub-optimal, back-translated data in general are noisy.

Considering the role of DA techniques in the literature, it is evident that the success of LRNMT increasingly depends on improving these techniques to produce *high-quality* parallel data. It underscores the necessity of coupling automatically generated parallel data with PDC strategies to ensure the creation of *high-quality* corpora. In our research, we explore the first three DA strategies and discuss them in detail in the subsequent chapters. We do consider the back-translation technique, as we believe the improvements in PDC can be applied to back-translated data to extract *high-quality* parallel sentences. The language pairs covered in this study include LRL-pairs, such as Sinhala-English (Epaliyana et al., 2021), English-Romanian (Caswell et al., 2019).

## 2.6 Sinhala-Tamil-English Related NMT

We conduct this research to improve the NMT between Sinhala-Tamil-English languages. The rationale for selecting these languages, and also the existing work conducted related to these languages, is discussed in this section. Finally, we describe the datasets which are available for the progression of research between the language pairs.

### 2.6.1 Selection of Low-Resource Language Pairs

Sri Lanka is a multi-ethnic country, in which `Sinhala`, `Tamil` are treated as National Languages in Sri Lanka. `Sinhala` (de Silva, 2025) belongs to the Indo-Aryan language family and is the native language of the largest ethnic group in Sri Lanka, which is about 16 Million people according to statistics in 2021. [1] `Tamil`, belongs to the Dravidian language family, and is spoken by the second largest ethnic group of about 5 Million people in Sri Lanka and more than 80 Million people (Jain et al., 2020) globally. In the context of Sri Lanka, only a very small population can communicate in both languages and hence `English` is considered the link language (de Silva, 2023). Therefore, to ensure effective communication, information is disseminated in all three languages. Which means all the official documents (Farhath et al., 2018b) such as annual reports, gazettes and circulars, would need to be published in all three languages. Currently, the translation task is conducted by humans, and hence, the exercise is labour intensive and inefficient. Therefore, reliable translation systems to translate between `Sinhala-Tamil-English` languages are crucial.

`Sinhala` exhibits complex morphological structures, including rich inflectional and derivational processes, but is classified as a LRL (Ranathunga and de Silva, 2022) due to the scarcity of linguistic resources and tools. Unlike `Sinhala`, `Tamil` benefits from a relatively larger digital presence, but it still faces challenges in NLP applications

[1] https://www.ethnologue.com/language/sin/

due to morphological complexity, agglutinative grammar, and resource limitations in certain domains. Further, Sinhala and Tamil belong to two distinct language families. Therefore when it comes to the non-English centric language-pairs we can generalize our findings by considering Sinhala-Tamil language pair. For Sinhala and Tamil languages, linguistic processing tools and resources such as morphological analysers, dependency parsers, or annotated datasets are scarce (de Silva, 2019). As a result, Sinhala and Tamil related NMT research still lags in achieving state-of-the-art results, compared to HRLs.

### 2.6.2 Progression NMT Research among Sinhala-Tamil-English Languages

Initial research on Sinhala related NMT was between the Sinhala-Tamil language pair by Tennage et al. (2017). This was improved with transliteration and Byte-pair-Encoding (BPE) (Tennage et al., 2018a). Afterwards, the transformer architecture with BPE has shown significant gains for the Sinhala-Tamil (Pramodya et al., 2020), English-Sinhala (Naranpanawa et al., 2020; Fonseka et al., 2020) and English-Tamil (Ramesh et al., 2021a) language pairs. Further improvements were obtained with BPE with back-translation for English-Sinhala (Nissanka et al., 2020; Pushpananda, 2019) and English-Tamil (Jain et al., 2020) language pairs. Subsequent work considered morphological segmentation driven-vocabulary (Krupakar and Milton, 2016), memory-efficient linguistic-driven vocabulary (Dhar et al., 2021) as an alternative to BPE and reported improvements. Several other works have focused on improving Named Entity Translation (NET), adapting statistical methods (Priyadarshani et al., 2019) and with denoising methods (Ranathunga et al., 2024b).

Thillainathan et al. (2021) explored the success of transfer learning by fine-tuning the mBART pre-trained model to improve the benchmark scores in the directions of English-Sinhala, English-Tamil and Sinhala-Tamildirections and vice versa. Pramodya (2023) explored mT5 (Xue et al., 2021a), for Sinhala-Tamil language-pair and show that BLUE score gains can be obtained upto +3.28. aSu et al. (2024) evaluated Parameter Efficient Fine-Tuning Techniques (PEFT) covering Sinhala-Tamil language-pair, and recommends 6 PEFT architectues favourable for LRLs. Sequence-to-sequence trained multiPLMs such as NLLB (Costa-jussà et al., 2022), mT5 (Xue et al., 2021a), M2M-100 (Fan et al., 2020) benefit the translation among Sinhala, Tamil and English.

These studies demonstrate the effectiveness of multilingual pre-trained models and fine-tuning strategies for improving translation quality among Sinhala, Tamil, and English. Despite these advances, challenges remain due to limited data, complex morphology in the languages.

### 2.6.3 Dataset Availability

SiTa-Trilingual dataset (Fernando et al., 2020) is a human-curated parallel dataset with 100K parallel sentences for `English-Sinhala`, `English-Tamil`and `Sinhala-Tamil` language-pairs. This is the only gold-standard dataset available for these language pairs.

The OPUS[2] collection contains parallel data supporting the three language pairs; however, the available OPUS parallel data are web-crawled and are noisy (Ranathunga et al., 2024a; Kreutzer et al., 2022b). Therefore, these parallel datasets need to be curated before using them or training NMT systems. Therefore, except for the SiTa-Trilingual government domain 100K dataset, there is a limitation of curated parallel datasets aimed at improving NMT between `English-Sinhala`, `English-Tamil` and `Sinhala-Tamil` language pairs.

## 2.7 Chapter Summary

This chapter outlined the background work related to MT and the advancements in terms of architecture and techniques related to NMT. Taking the DA technique as a viable approach to improve LRNMT, we discussed the various strategies employed to generate or enhance parallel corpora, particularly for LRL pairs. We highlighted the significance of data quality and coverage, and how augmentation techniques such as synthetic data generation, back-translation, and filtering heuristics contribute to improved translation performance. This forms the foundation for the experimental investigations presented in the subsequent chapters.

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

# CHAPTER 3

# GENERATING SYNTHETIC PARALLEL SENTENCES

## 3.1 Introduction

In this chapter, we present the first data augmentation strategy we have explored: word or phrase replacement-based augmentation strategy. Our approach focuses on generating synthetic parallel sentence pairs in a Low-Resource Language (LRL) setting.

Existing work taking the direction of word/phrase replacement-based augmentation imposes limited mechanisms for validating the final synthetic sentence-pair in terms of syntactic and semantic correctness. Word/phrase replacement was introduced by Fadaee et al. (2017), where only the replacement context was validated by means of a Language model score. This work was improved by Tennage et al. (2018b), where they validated the replacement by means of Part-of-Speech (POS) and morphological agreement. On the otherhand (Peng et al., 2020) limits to validating the replacement considering semantic simlarity. In contrast to previous research, our Data Augmentation (DA) technique incorporates both syntactic and semantic constraints to ensure the plausibility of the synthetic sentences. As syntactic constraints, we use POS and morphological information similar to Tennage et al. (2018b) and as semantic constraints, we use word embedding-based semantic similarity similar to Peng et al. (2020). Additionally, we propose selecting the candidate sentences for replacement using a sentence similarity score.

As the words to be replaced, we consider augmenting Out-of-Vocabulary (OOV) terms, which is a hindrance for the NMT models to provide reliable translations. Our intuition is that the proposed approach benefits NMT for LRLs in two key ways. First, augmenting rare and unseen words in the training dataset—typically treated as OOV helps to improve translation quality for sequences containing such words. Secondly, the augmented dataset leads to improving the available existing training data for the language-pair. The rare word and dictionary term augmentation is described in detail in Section 3.3. Once the synthetic sentences have been generated, they are combined with the existing training dataset and the NMT model is trained. We conduct this work for the English-Sinhala language pair. The maximum gains are from rare word augmentation, +0.91 and +0.74 for English-Sinhala and Sinhala-English directions. For dictionary term augmentation, the maximum gains reported are +0.74 and +0.18 for English-Sinhala and Sinhala-English directions.

With this augmentation strategy, we address the first research objective, **RO1. Propose and implement an algorithm to generate synthetic parallel sentences to augment out-of-vocabulary terms**.

## 3.2 Related Work

Recent research on word or phrase replacement-based techniques for NMT focuses on different linguistic levels when inducing synthetic parallel sentences. Word-level augmentation aims at substituting words in existing parallel sentences to provide diverse contexts for the replaced word (Wang et al., 2018b; Fadaee et al., 2017). As the words for augmenting, mainly rare-words (Tennage et al., 2018b; Fadaee et al., 2017), which are words with low occurrence and words from a lexicons (Wang et al., 2018b) have been considered. Phrase and subtree-level methods introduce syntactic variation through phrase substitutions (Alam et al., 2024; Liu et al., 2023) or dependency subtree (Nagy et al., 2023) manipulations. Sentence-level augmentation generates rephrased sentences to enhance lexical diversity and improve translation fluency (Maimaiti et al., 2022; Gao et al., 2023). However, the success of these methods rely on the accuracy of the linguistic tools to extract word-level, phrase-level or sub-tree-level information.

Fadaee et al. (2017) is the first to explicitly focus on a word replacement method to augment the rare words. In this technique, for a considered parallel sentence pair, a common word in the source sentence is replaced by a rare word in the source language. The synthetic target side sentence is obtained by replacing the aligned target side common word with the translation of the source rare word. As the synthetically generated sentences lacked fluency, Tennage et al. (2018b) incorporated linguistic constraints to validate the replacement. Here, the word was replaced if the rare word and the common word identified in the sentence agreed in terms of POS and morphology. In a similar work, Duan et al. (2020) relied on dependency information to determine the suitable word to be replaced in the sentence. In both these techniques, although the syntactic correctness of the synthetic sentence was preserved, semantic correctness could not be guaranteed.

Another common approach is to augment terms from a bilingual lexicon to produce synthetic parallel sentences. Nag et al. (2020) used a bilingual dictionary to translate target-side monolingual sentences word-for-word in order to generate source-side synthetic sentences. However, for morphologically rich languages, this method often yields sub-optimal results, as dictionary entries are typically in their base forms and do not account for inflected variants. Alternatively, Peng et al. (2020) employed an in-domain dictionary to induce synthetic parallel sentences from an out-of-domain parallel corpus using phrase-replacement augmentation based solely on semantic similarity. They first filtered the topmost semantically similar sentences by comparing the source-side dictionary terms with source-side sentences from the out-of-domain corpus. Then, they identified the noun phrase to replace in the candidate sentence by measuring the semantic similarity between the dictionary term and the noun phrases within the sentence. Nevertheless, the approach remains sub-optimal, as it does not ensure syntactically correct replacements.

## 3.3 Methodology

Our DA approach adopts a word/phrase replacement-based strategy that leverages both syntactic and semantic constraints to generate synthetic parallel sentences. Overcoming the limitations in the existing work, we impose validations to ensure **both** syntactic and semantic plausibility of the final synthetic sentence-pair. As syntactic constraints, we use POS and morphological agreement and as semantic constraints, we use the semantic similarity at the word-level and sentence-level, prior to replacement. We discuss this in detail in this section. The augmentation is two-fold:

**Rare word augmentation:** For a considered sentence pair in the existing training corpus, a candidate word in the source-side sentence is replaced with a rare word identified from the source side of the parallel corpus, ensuring that both syntactic and semantic constraints are satisfied. This results in the generation of a synthetic source-side sentence. In a similar manner, the synthetic target-side sentence is created by replacing the aligned word or phrase from the target sentence with the translation of the rare word.

**Dictionary augmentation:** To generate a synthetic source-side sentence from a given parallel sentence pair, a selected word in the source-side sentence is replaced with a source-side dictionary term that satisfies the same syntactic and semantic constraints. The corresponding target-side synthetic sentence is then created by replacing the aligned word or phrase with the appropriate target-side dictionary term.

### 3.3.1 Rare Word Augmentation

In this data augmentation technique, rare words are substituted into existing parallel sentences to introduce novel contexts. The rare word augmentation process is illustrated in Figure 3.1, and a step-by-step description is provided in Sections 3.3.1.1 to 3.3.1.6.

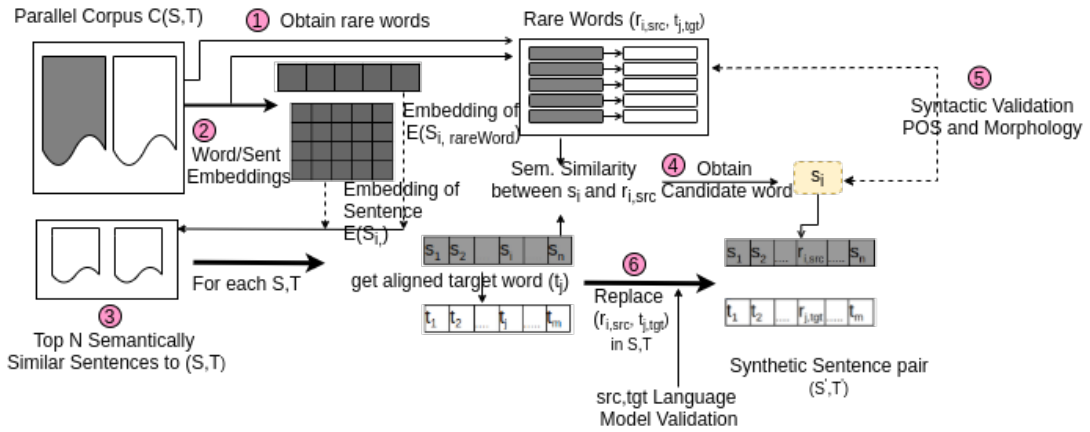


Figure 3.1: Data Augmentation Process.

#### 3.3.1.1 Obtain Rare words

We identify rare words from the source side of the parallel corpus following the approach of Fadaee et al. (2017) and Tennage et al. (2017). Words with a frequency below a specified threshold ($T_R$) are considered rare. A word-alignment model is then trained on the existing parallel sentences to obtain the corresponding target-side rare words. This process is depicted as Step 1 in Figure 3.1.

#### 3.3.1.2 Word/Sentence Embeddings:

Pre-trained word embeddings are effective in capturing both linguistic and word-level constraints, a property that has been empirically demonstrated for Sinhala as well (Lakmal et al., 2020). However, their potential has not yet been explored for data augmentation in the context of Sinhala. In our work, we incorporate semantic information in two key steps: (1) to filter candidate sentences from the existing parallel corpus for rare word replacement, as described in Section 3.3.1.3; and (2) to identify the specific word to be replaced within those sentences, as outlined in Section 3.3.1.4.

Although embeddings are widely used in Natural Language Processing (NLP) tasks, different types of embeddings capture word-level constraints in distinct ways. This distinction becomes particularly evident when retrieving the most similar words (Mikolov et al., 2013). For instance, some embeddings are more effective at capturing syntactic similarities between words (e.g., run – running), while others are better suited for capturing semantic similarities (e.g., sing – chant).

Artetxe et al. (2018b) empirically proved that applying a linear transformation to embedding vectors as a post-processing step enables them to capture the desired type of similarity, either semantic or syntactic similarity between words. To optimize the embeddings for the DA task with an emphasis on semantic similarity, we adopt this approach by applying a post-processing transformation parameterized by a value $\alpha$. As the optimal $\alpha$ must be determined empirically, we evaluate the performance of word embeddings transformed with various $\alpha$ values in DA experiments. The $\alpha$ value that yields the highest performance is then selected for post-processing the embeddings in all subsequent experiments. Sentence embeddings were computed by averaging the post-processed embeddings of the individual words in the sentence. This corresponds to Step 2 in Figure 3.1.

#### 3.3.1.3 Obtaining Candidate sentences

Identifying the most suitable sentences for rare word replacement is a critical step. We first compute the semantic similarity between the source-side training sentences and the sentence containing the source-side rare word. The most similar sentences are then selected as candidates for data augmentation. This process corresponds to Step 3 in

Figure 3.1.

#### 3.3.1.4 Obtaining the Candidate Word

In Step 4 of Figure 3.1, the word to be replaced is identified by computing the cosine similarity between the source-side rare word and each word in the source sentence. The word with the highest cosine similarity is selected as the candidate for replacement.

#### 3.3.1.5 Syntactic Validation

In Step 5, the identified candidate word is further evaluated for syntactic agreement with the rare word. Following the approach of Tennage et al. (2018b), we check POS and morphology agreement for Sinhala words. For English words, we consider only number agreement. Due to the lack of a syntactic parser for Sinhala, further syntactic constraints, such as dependency rules (Duan et al., 2020), were not considered.

#### 3.3.1.6 Generating Synthetic Sentences

In Step 6, the candidate word identified in Step 5 is replaced with the source-side rare word to generate the synthetic source sentence.

A word-alignment model is then used to identify the corresponding word or phrase in the target-side sentence that should be replaced. This identified unit is substituted with the target-side rare word or phrase to produce the synthetic target sentence.

Subsequently, the replacement context is validated using a Language Model (LM) trained on monolingual corpora in the respective languages. The source-side context is scored using the source-side LM, while the target-side trigram context is scored using the target-side LM.

The replacement is accepted if the ratio between the LM score of the synthetic sentence and that of the original sentence exceeds a specified threshold on both the source and target sides.

### 3.3.2 Dictionary Augmentation

The dictionary augmentation algorithm closely mirrors the rare word augmentation algorithm, with the exception of Step 3 in Figure 3.1. As shown in the DA results in Table 3.4, sentence filtration did not yield any improvement; therefore, this step was omitted in the dictionary augmentation process. Consequently, all source-side sentences has been considered as candidate sentences for dictionary term replacement. As with the rare word augmentation, the embeddings were post-processed prior to their use in the DA experiments. From Step 4 onwards, the dictionary augmentation process follows the same sequence of steps as the rare word augmentation approach, as illustrated in Figure 3.1.

### 3.3.3 Combined Solution

Finally, we conduct a combined experiment by merging the synthetic parallel data sets, augmenting rare words and dictionary terms. The combined augmented data is added on top of the 56k training set, and we train the NMT model.

## 3.4 Experiments

In this section, we describe the datasets used (Section 3.4.1) and the experiments conducted (Section 3.4.2) in this research work. First, we define three baseline models as described in Section 3.4.3. Then we train comparison baseline models to evaluate whether merely duplicating sentences or random replacement yields gains. Finally, we conduct an ablation study to analyse the effectiveness of using a single constraint and combinations of them in the augmentation. The NMT models are trained by combining the augmented parallel sentences with the existing seed training dataset.

### 3.4.1 Dataset

We use the `Sinhala-English` human-curated parallel sentences from the SiTa-Trilingual parallel corpus (Fernando et al., 2020) from the government document domain in our experiments. However, the final statistics of this dataset are not available to disclose. This is the only human-curated dataset available between the En-Si language pair. Alternative parallel English-Sinhala datasets are available on OPUS[1]. These datasets have been reported to be noisy (Ranathunga et al., 2024a; Bane et al., 2022). Hence, those corpora were not considered in this work. The corpus statistics are given in Table 3.1.

Table 3.1: Parallel Corpus Statistics of Training and Validation sets

| | **Train** | **Validation** |
|---|---|---|
| No. Sentences | 54914 | 1623 |
| No. Words (`En`) | 553002 | 23578 |
| No. Words (`Si`) | 535185 | 22721 |

The test set statistics are given in Table 3.2. TS1 is the evaluation set from the SITA government domain dataset. We have sampled different evaluation sets (TS2 and TS3), containing different numbers of rare words and dictionary terms, to analyse the effectiveness of our approach. These were also from the government domain. Here, Dic. Terms (OOV) refer to the number of dictionary terms in the test sets that do not appear in the training data.

We obtain `Sinhala` and `English` monolingual data (Isuranga et al., 2020) to train the language models in the respective languages. The corpus is a combination of

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

Table 3.2: Test set Statistics

| | TS1 | TS2 | TS3 |
|---|---|---|---|
| No. Sentences | 1603 | 1462 | 1438 |
| **Sinhala** | | | |
| No. Words | 18513 | 28918 | 26308 |
| Unique Words | 4520 | 5341 | 5057 |
| Rare Words | 76 | 133 | 127 |
| Dic. Terms | 502 | 596 | 594 |
| Dic. Terms (OOV) | 11 | 17 | 23 |
| **English** | | | |
| No. Words | 19248 | 30437 | 27815 |
| Unique Words | 4237 | 4956 | 4865 |
| Rare Words | 55 | 55 | 68 |
| Dic. Terms | 1314 | 1804 | 1739 |
| Dic. Terms (OOV) | 58 | 108 | 99 |

publicly available government data, common crawl and news data. The statistics are detailed in Table 3.3. We use the SRILM (Stocke, 2011) toolkit to train a tri-gram LM, similar to Fadaee et al. (2017).

Table 3.3: Statistics corresponding to the Monolingual Corpus to train the LMs.

| | English | Sinhala |
|---|---|---|
| No of Sentences | 1,286,945 | 1,163,675 |
| No of Words | 51,193,388 | 48,283,636 |

For the dictionary augmentation, we use an `English-Sinhala` dictionary[2] extracted from public data with a total of 23660 terms.

### 3.4.2 NMT Experiment Setup

We used the encoder–decoder architecture with attention for our NMT experiments, following Bahdanau et al. (2015). However, the proposed DA technique is architecture-independent and can be applied to train with any sequence-to-sequence model.

We train the NMT models using the OpenNMT toolkit (Klein et al., 2018). The experiments have been conducted on Google Colaboratory using an NVIDIA K80 GPU with 8GB of RAM. The NMT encoder is a two-layer bidirectional Long Short-Term Memory (LSTM) network, and the decoder is a two-layer LSTM using global attention (Bahdanau et al., 2015). We train the models with a batch size of 32, a dropout rate of 0.4, and using the Adam optimizer, following the work of Epaliyana et al. (2021).

[2] `https://www.maduraonline.com/`

We conduct a pre-tokenisation to separate the punctuations and the words using the Moses tokenizer (Koehn et al., 2007) for English and a custom-built tokenizer (Farhath et al., 2018a) for Sinhala. Subsequently, we perform BPE tokenization to generate a vocabulary of 25,000 tokens for training the BiLSTM NMT model. We report our results using the BLEU score metric on the three test sets using the multi-bleu.perl script (Papineni et al., 2002). To account for training variance, each experiment has been conducted three times, and the average BLEU score is reported.

Finally, we conduct these experiments on vanilla transformers (Vaswani et al., 2017) sequence-to-sequence architecture and analyse the impact on our approach. We first train a Sentencepiece[3] tokenization model with 25,000 using the training dataset. Here we run the experiment only with the augmentation set, which yielded the best NMT gains for rare-word and dictionary term augmentation. We report the results using BLEU score (Popović, 2017) metric for comparison purposes.

### 3.4.3 Baseline Models

We create three baseline experiments as follows:

- **Baseline[train54K]:** Our initial baseline model is trained with 54K parallel sentences without augmentation.
- **Fadaee et al. (2017):** Augmented parallel sentences are generated using the technique proposed by Fadaee et al. (2017). These sentences, along with the original 54K training instances, has been used to train the NMT model. This setup serves as our second baseline.
- **Peng et al. (2020):** An augmented parallel sentence is generated using the technique proposed by Peng et al. (2020). The original 54K training instances, together with the augmented sentences produced by this method, are used to train the third baseline NMT model.

For comparison with the baseline, NMT models has been trained by randomly replacing rare words and dictionary terms in existing sentences, without considering any syntactic or semantic constraints. We conduct a second comparison experiment by duplicating random samples of 10K, 25K, and 35K sentences from the training data. These experiments were designed to evaluate whether simple duplication alone could lead to performance improvements. The results of these experiments are presented in Table 3.4.

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

### 3.4.4 Augmentation of Rare Words

Following Tennage et al. (2017), words with a frequency threshold $T_R = 1$ in the training corpus has been considered as rare words. A total of 3133 and 2370 valid rare words could be identified from the Sinhala and English sides, respectively. GIZA++ (Och and Ney, 2003) automatic word alignment algorithm is used to determine the translation word or phrase from the parallel target side sentence.

To obtain POS and morphology information for Sinhala, the TnT POS Tagger (Fernando and Ranathunga, 2018) and sin-morphy (Kumarasinghe et al., 2021) were used, respectively. For the English side, the Python NLP library Spacy[4] was used. The tri-gram LM threshold has been chosen as 0.6 for context validation.

We used fastText embeddings (Bojanowski et al., 2016) to obtain word representations for Sinhala and English. However, fastText tends to return syntactically similar words rather than semantically similar ones. Since our objective was to identify the semantically similar words, we applied the post-processing method proposed by Artetxe et al. (2018b) to enhance the embeddings.

To determine suitable $\alpha$ values for post-processing, we conducted the dictionary augmentation experiment using word similarity alone, based on post-processed embeddings by varying $\alpha$ values. Empirically, we set $\alpha$ to $-0.15$ for Sinhala word embeddings and $0.15$ for English word embeddings, respectively. As shown in Table 3.4, the rows Baseline+wordSim$_{w/o\ pp}$ and Baseline+wordSim demonstrate that applying post-processing with these identified $\alpha$ values improves BLEU scores, with gains of up to +0.63 and +0.60 in the respective translation directions.

These augmentation experiments were conducted on Google Colaboratory on an NVIDIA K80 GPU with 8GB VRAM.

An ablation study has been conducted by incorporating (1) syntactic constraints only, (2) semantic constraints only, and (3) a combination of both syntactic and semantic features in the DA experiments. The objective is to identify the most effective feature configuration for the augmentation task. The results of this study are presented in Table 3.4.

### 3.4.5 Augmentation of Dictionary

The Sinhala entries in the dictionary predominantly consisted of phrases. For these multi-word terms, embeddings were computed by averaging the individual fastText word embeddings. The resulting dictionary term embeddings were then post-processed using the previously identified $\alpha$ values.

The dictionary augmentation experiments were carried out using (1) syntactic constraints only, (2) semantic constraints only, and (3) a combination of both. The

[4] https://spacy.io/

results are presented in Table 3.5.

### 3.4.6 Combined Experiments

As a combined experiment, the pseudo-parallel datasets that yielded the best performance in the rare word augmentation and dictionary augmentation experiments has been merged with the original parallel corpus to train the NMT model. The results of this experiment are presented in Table 3.5.

## 3.5 Results Analysis

The best NMT scores obtained from rare-word augmentation and dictionary term augmentation surpassed the re-created baselines of Fadaee et al. (2017) and Peng et al. (2020), demonstrating that our data augmentation, incorporating both syntactic and semantic constraints, results in high-quality parallel data, resulting in better gains.

In contrast, random replacements yielded only marginal improvements. In the duplication experiments, randomly duplicating 25K training samples led to BLEU score increases of +0.23 and +0.41 in the respective translation directions. However, further increasing the duplicated sample size resulted in performance degradation, indicating that mere duplication is less effective compared to a data augmentation technique.

Similar experiments were also conducted using dictionary terms. As shown in Table 3.4, this form of random augmentation did not lead to any improvements and, in fact, resulted in a reduction from the baseline scores.

### 3.5.1 Rare Word Augmentation

According to the results in Table 3.4, the comparable performance between the *Baseline+wordSim* and *Baseline+pos+morph* experiments confirms that, for LRLs, the use of embeddings is promising. Furthermore, we observe that combining semantic and syntactic constraints consistently improves results in both translation directions.

We initially expected the *pos+morph* experiments to outperform those using only *pos*. However, in the Si→En direction, a decrease in performance was observed. This can be attributed to the reduced augmented dataset size. We suspect that Sinhala morphological parser's limited coverage, failing to produce morphological information for some rare words. In contrast, in the En→Si direction, *pos+morph* experiments showed improvements over *pos*-only setups. With rare word augmentation, we achieved the highest BLEU score gains of +0.91 in the Si→En direction and +0.74 in the En→Si direction.

Table 3.4: Rare Word Augmentation Results considering different syntactic and semantic constraints

| Experiment | Aug. Sent. | Si → En (BLEU) | | | Aug. Sent. | En → Si (BLEU) | | |
|---|---|---|---|---|---|---|---|---|
| | | TS1 | TS2 | TS3 | | TS1 | TS2 | TS3 |
| Baseline [train54K] | - | 22.47 | 21.22 | 26.82 | - | 20.61 | 19.33 | 24.97 |
| Baseline (Fadaee et al., 2017) | 10947 | 22.76 | 21.28 | 26.89 | 13675 | 20.80 | 18.95 | 24.62 |
| Baseline (Peng et al., 2020) | 12447 | 22.63 | 21.06 | 26.62 | 1215 | 20.49 | 19.30 | 25.37 |
| **Random Duplicating** | | | | | | | | |
| Baseline+randDuplicate10K | 10000 | 22.40 | 20.89 | 26.30 | 10000 | 20.39 | 19.12 | 24.48 |
| Baseline+randDuplicate25K | 25000 | 22.65 | 21.29 | 27.05 | 25000 | 21.00 | 19.44 | 25.38 |
| Baseline+randDuplicate35K | 35000 | 22.59 | 21.05 | 26.76 | 35000 | 20.25 | 19.38 | 25.33 |
| **Random Replacement** | | | | | | | | |
| Baseline+randRareWords10K | 10000 | 22.26 | 20.53 | 26.25 | 10000 | 20.67 | 19.33 | 25.11 |
| Baseline+randDictionary10K | 10000 | 22.50 | 20.77 | 26.56 | 10000 | 20.61 | 18.60 | 24.60 |
| **Linguistic Constraints** | | | | | | | | |
| Baseline+pos | 2276 | 22.56 | 21.44 | 27.46 | 2587 | 20.76 | 19.44 | 25.33 |
| Baseline+pos+morph | 1560 | 22.40 | 21.50 | 27.43 | 2760 | 20.99 | 19.33 | 25.35 |
| **Word Similarity** | | | | | | | | |
| Baseline+wordSim$_{\text{wo pp}}$ | 8684 | 22.18 | 21.23 | 26.65 | 7792 | 20.48 | 18.78 | 25.08 |
| Baseline+wordSim | 7667 | 22.35 | 21.39 | 27.28 | 7544 | 21.08 | 19.23 | 25.12 |
| Baseline+wordSim+pos | 1789 | **22.88** | **21.84** | **27.73** | 3780 | 20.88 | **19.51** | 25.56 |
| Baseline+wordSim+pos+morph | 927 | 22.34 | 21.47 | 27.55 | 1780 | 20.89 | 19.47 | 25.53 |
| **Word Similarity + Sentence Similarity** | | | | | | | | |
| Baseline+wordSim+sentSim | 7518 | 22.57 | 21.40 | 27.11 | 6642 | 20.97 | 19.07 | 25.13 |
| Baseline+wordSim+sentSim+pos+morph | 854 | 22.42 | 21.56 | 27.64 | 130 | **21.18** | 19.40 | **25.71** |

### 3.5.2 Dictionary Augmentation

The results for dictionary augmentation, presented in Table 3.5, show a reduction in BLEU scores for the Si→En direction. However, performance improved progressively with the incorporation of syntactic and semantic constraints. Despite the initial drop, our results still surpass the recreated baselines of Fadaee et al. (2017) and Peng et al. (2020). From the dictionary augmentation experiments, the highest BLEU score gain of +0.71 was achieved in the En→Si direction.

It is observed that dictionary-augmented datasets containing more than 12K synthetic sentences consistently resulted in BLEU scores lower than the baseline, in both translation directions. Notably, the only dictionary augmentation experiment *Baseline+wordSim+pos+morph* that generated approximately 7K sentences yield the highest gain of +0.71 BLEU points.

A similar trend is observed in the rare-word augmentation experiments, where performance declined when the number of synthetic sentences exceeded 8K. These findings indicate a potential negative impact associated with incorporating large volumes of synthetic data, particularly when the augmented set surpasses 12K sentences.

In the combined experiments, the parallel augmented sentences were 19K and 7K in

Table 3.5: Dictionary Word Augmentation Results considering different syntactic and semantic constraints. Here, TS1, TS2 and TS3 correspond to the three evaluation sets.

| Experiment | Aug. Sent. | Si → En (BLEU) | | | Aug. Sent. | En → Si (BLEU) | | |
|---|---|---|---|---|---|---|---|---|
| | | TS1 | TS2 | TS3 | | TS1 | TS2 | TS3 |
| Baseline [train54K] | - | 22.47 | 21.22 | 26.82 | - | 20.61 | 19.33 | 24.97 |
| Baseline (Fadaee et al., 2017) | 35901 | 21.59 | 19.36 | 22.70 | 49211 | 20.31 | 17.59 | 22.39 |
| Baseline (Peng et al., 2020) | 4856 | 22.28 | 20.76 | 26.17 | 5709 | 20.85 | 19.24 | 24.75 |
| **Linguistic Constraints** | | | | | | | | |
| Baseline+pos | 26940 | 22.37 | 20.84 | 25.49 | 15201 | 20.63 | 18.41 | 24.15 |
| Baseline+pos+morph | 18770 | **22.65** | 21.25 | 26.38 | 15201 | 20.50 | 18.76 | 24.26 |
| **Word Similarity** | | | | | | | | |
| Baseline+wordSim | 32170 | 21.57 | 20.39 | 24.96 | 57288 | 19.95 | 18.20 | 22.04 |
| Baseline+wordSim+pos | 18209 | 21.51 | **21.29** | **26.40** | 25651 | 20.26 | 18.52 | 23.64 |
| Baseline+wordSim+pos+morph | 12594 | 22.07 | 20.87 | 26.21 | 6721 | **21.02** | **19.42** | **25.68** |
| **Combined Experiment** | | | | | | | | |
| Baseline+rareWord+dicTerm | 19998 | 22.17 | 20.69 | 26.13 | 7031 | **20.66** | 19.31 | **25.55** |

the Si→En and En→Si directions respectively. Consistent with the previous behaviour, the BLEU score gain is observed only in En→Si direction as +0.58. However, we need to conduct more experiments to identify the optimal ratio between the synthetic sentences and the parallel corpus to achieve the best scores.

Considering the number of rare words and dictionary terms present in the test sets as in Table 3.2, we observe that even if the OOV terms are present in low counts, the DA technique is still effective enough to improve the overall BLEU score. This is mainly owing to the improvement in the overall translation and its fluency. This is further evident in the example in Table 3.6.

### 3.5.3 Qualitative Analysis

To analyse our second contribution, whether the OOV augmentation improves the translation output, we consider a *Sinhala* sentence containing a rare word and observe how the translation changes with the DA experiment. As shown in Table 3.6, the selected sentence has the rare word ie. පරිශීලනය (reference). We observe that the correct translation of the rare word is generated when using only syntactic or only semantic constraints. However, a more fluent output is generated when both constraints are combined. Therefore, it is evident that DA aids the NMT to improve the translation output for the sequences containing OOV terms.

### 3.5.4 NMT Results on Transformer Architecture

The NMT results for the transformer architecture are shown in Table 3.7. The best NMT result is yielded when both syntactic and semantic constraints are considered in the aug-

Table 3.6: Improvement in the En translation with respective to each augmented dataset. The input Si sentence contains *parisilanaya*, the OOV term. Using more syntactic and semantic constraints improves the fluency and completeness of the translated sentence.

| | |
|---|---|
| Rare word | පරිශීලනය (parisílanaya) |
| Si Sentence | විනිශ්චයකාරවරුන්ගේ පරිශීලනය පිණිස පුස්තකාලය සඳහා 'නීතිය' පිළිබඳ නව ග්‍රන්ථ මිල දී ගන්නා ලදි.<br>*viniścayakāravarungē **parišílanaya** piṇisa pustakālaya sañdahā 'nītiya' piḷibañda nava grantha mila dī gannā ladi.* |
| En Sentence (Ref.) | New books on "Law" were purchased for the library for the **reference** of the judges. |
| Baseline [train54K] | new law for the library for the library for the library was purchased. |
| Baseline+pos+morph | new law Books were purchased for the Library **reference** to the Judges. |
| Baseline+wordSim | new law Books were purchased on the Library **reference** for the Library reference. |
| Baseline+wordSim+pos | new law Books were purchased for the Library for easy **reference** of the Judges. |

mentation, with the exception of testsets TS2 in the En→Si direction. However, in this situation, the *Baseline+wordSim* is almost comparable to *Baseline+wordSim+pos*. For dictionary term augmentation also considering both syntactic and semantic constraints as a combination produce the best gain. Therefore, it is safe to say that our hypothesis is further justified with the transformer architecture as well.

Another observation is that the performance gains obtained using the transformer architecture are greater than those achieved with the LSTM-based sequence-to-sequence architecture. We believe the transformer's ability to capture long-range dependencies and attend to relevant contextual signals more effectively resulted in these improvements in gains.

The combined experiments did not show superior gains compared to the best score obtained with rare word and dictionary term augmentation for the majority of the cases. However, with the exception of testsets TS2 and TS3 in the En→Si direction. This is likely due to the same limitation seen with the LSTM model. Therefore, more experiments would be needed to evaluate an optimal training set to augmented set ratio, for our approach to be favourable to improve NMT results further.

## 3.6 Discussion

Although high-accuracy linguistic resources are available for obtaining syntactic information for English, this is not the case for the Sinhala language. The Sinhala PoS Tagger is an SVM classifier, while the Sinhala morphological analyser was implemented using rules. As a result, morphological information was not available for all the words in the training corpus. This limitation impacted our augmentation strategy in two ways. Firstly, the limited vocabulary coverage of the morphological analyser restricted the number of words that could be considered for replacement. Secondly, it reduced the number of synthetic sentences generated.

Further, we rely on statistical models for word alignment and language modelling.

Table 3.7: NMT Results on transformer architecture, trained with the best performing syntactic and semantic combination from the rare word and dictionary term augmentation experiments.

| Experiment | Si → En (BLEU) | | | Experiment | En → Si (BLEU) | | |
|---|---|---|---|---|---|---|---|
| | TS1 | TS2 | TS3 | | TS1 | TS2 | TS3 |
| **RareWords Augmentation** | | | | | | | |
| Baseline | 21.19 | 15.63 | 16.16 | Basleline | 20.19 | 13.41 | 16.23 |
| Baseline (Fadaee et al., 2017) | 22.18 | 16.80 | 18.69 | Baseline (Fadaee et al., 2017) | 20.99 | 13.58 | 16.52 |
| Baseline (Peng et al., 2020) | 23.14 | 18.13 | 19.25 | Baseline (Peng et al., 2020) | 20.64 | 14.18 | 17.13 |
| Baseline+pos+morph | 23.19 | 17.73 | 19.92 | Baseline+pos+morph | 20.63 | 14.04 | 17.21 |
| Baseline+wordSim | 22.66 | 18.12 | 19.87 | Baseline+wordSim | 20.73 | **14.47** | 17.34 |
| Baseline+wordSim+pos | **23.27** | **18.49** | **20.15** | Baseline+wordSim+pos | **21.41** | 14.40 | **17.38** |
| **Dictionary Augmentation** | | | | | | | |
| Baseline | 21.19 | 15.63 | 16.16 | Basleline | 20.19 | 13.41 | 16.23 |
| Baseline (Fadaee et al., 2017) | 20.02 | 16.01 | 16.50 | Baseline (Fadaee et al., 2017) | 19.54 | 13.04 | 15.48 |
| Baseline (Peng et al., 2020) | 20.60 | 15.30 | 17.06 | Baseline (Peng et al., 2020) | 20.57 | 13.32 | 15.80 |
| Baseline+pos+morph | 21.08 | 15.24 | 17.42 | Baseline+pos+morph | 20.43 | 14.56 | 17.45 |
| Baseline+wordSim+pos | **22.87** | **17.40** | **20.19** | Baseline+wordSim+pos | 19.42 | 13.15 | 16.40 |
| Baseline+wordSim+pos+morph | 21.66 | 16.64 | 17.76 | Baseline+wordSim+pos+morph | **21.52** | **14.85** | **18.77** |
| **Combined Augmentation** | | | | | | | |
| Baseline+rareWord+dicTerm | 23.10 | 17.11 | 20.06 | Baseline+rareWord+dicTerm | 21.23 | **15.34** | **19.01** |

The alignment model is used to identify the target-side word that needs to be replaced. The sub-optimal nature of this model can lead to incorrect word replacements on the target side, which might result in an ill-formed target side synthetic sentence. The tri-gram LMs may not capture complex syntactic and semantic patterns, which may lead to producing low scores for the replacement context, which may lead to rejecting such a synthetic sentence as a result of the language model's limitation.

In Figure 3.2 we show a limitation with the word-alignment model and the morphological analyser. Although the technique imposes mechanisms to produce a syntactically and semantically correct sentence, the suboptimal nature of the linguistic tools and models still produce errornous sentences as shown in the exmaple.

## 3.7 Chapter Summary

In this chapter, we addressed our first research objective, **RO1. Propose and implement an algorithm to generate synthetic parallel sentences to augment out-of-vocabulary terms.** Here we followed the word or phrase replacement-based augmentation strategy and proposed two algorithms to augment rare words and unseen words in the training corpus (using a bilingual dictionary), and to induce synthetic parallel sentences. We improved the existing work by imposing a combination of syntactic and semantic constraints to produce high-quality synthetic parallel sentences. We observed that

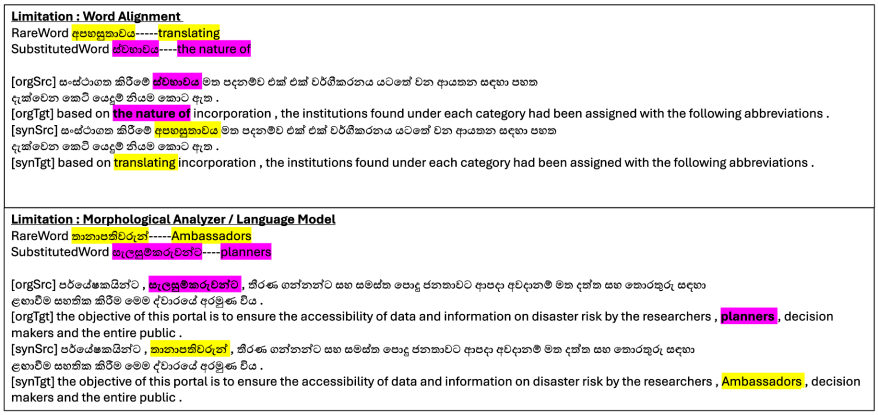


Figure 3.2: Shows the limitation with the word alignment (GIZA++) model and the limitation with the morphological analyser.

training the NMT model with these augmented synthetic parallel sentences leads to improved NMT scores and contributes to better translation outputs for sequences containing OOV terms. We further observed that using semantic constraints alone in the augmentation pipeline produced comparable results to what was produced by using syntactic constraints. We recommend employing semantic constraints alone in generating synthetic sentences as a viable approach for LRLs, which lack linguistic tool support (to obtain PoS or morphological information). Finally, we show that by combining both syntactic and semantic constraints, the results can be improved further.

# CHAPTER 4

# EMPIRICAL STUDY: MULTIPLMS FOR BITEXT MINING

## 4.1 Introduction

In the previous chapter, we explored the Data Augmentation (DA) technique of word or phrase replacement-based augmentation to induce synthetic parallel sentences to address the parallel data scarcity problem in Low Resource Languages (LRLs). In this chapter, we explore the most widely adopted method for acquiring parallel data, namely *bitext mining*, which aims to automatically extract parallel sentences from web-crawled *comparable corpora*. According to Burchell et al. (2025); El-Kishky et al. (2020); Bañón et al. (2020) the common pipeline of subtasks for *bitext mining* consists of web-crawling of monolingual data, document alignment and sentence alignment.

Document alignment refers to the subtask of identifying aligned web documents that contain *comparable corpora* (El-Kishky et al., 2020; Buck and Koehn, 2016a). The objective of sentence alignment is to find parallel sentence pairs in the already identified aligned document pairs. The recent techniques rely on vector representations (Costa-jussà et al., 2022) of the source side and target side documents or sentences and determine alignment using a semantic similarity measurement such as cosine similarity (Bañón et al., 2020). Therefore, the accuracy of both these subtasks are crucial towards producing *high-quality* parallel sentences, from the bitext mining pipeline.

Quality audits conducted by Ranathunga et al. (2024a); Bane et al. (2022); Kreutzer et al. (2022b) show that mined web corpora are noisy for LRLs. This is due to the weaker embeddings produced by the Multilingual Pre-trained Language Models (multiPLMs) to determine the alignment (Sloto et al., 2023) between the parallel data or the inherent noise (Moon et al., 2023) in the web corpora itself. To investigate the former, we conduct an empirical study and analyse what type of multiPLMs would produce optimal (or sub-optimal) results for the document alignment and sentence alignment tasks, to produce quality (or noisy) parallel sentences following the bitext-mining pipeline.

This study is aimed at answering the second research objective, **RO2: Conduct an empirical Study to determine the impact of different characteristics of the Pre-trained Multilingual Language Models on the Document Alignment and Sentence Alignment tasks for LRLs**. We experiment with three language pairs: `Sinhala-Tamil`, `Sinhala-English` and `Tamil-English`. First, we evaluate the impact of the embeddings obtained by three multiPLMs, LASER2
(Artetxe and Schwenk, 2019b), XLM-R (Conneau and Lample, 2019) and LaBSE (Feng et al., 2022) obtained embeddings on the selected sub-tasks. The rationale for selecting these multiPLMs is described in Section 4.3.2. Secondly, we evaluate whether the performance can be further enhanced by an improved scoring function, proposed by

Sachintha et al. (2021) using bilingual lexicons.

## 4.2 Related Work

In this study, we focus on the document alignment and sentence alignment subtasks in the bitext mining pipeline, which impact the quality of the induced parallel data.

### 4.2.1 Document Alignment

Automatic alignment of documents determines the probability that two documents are mutual translations of each other. Early methods in this area largely depended on metadata cues such as URLs (Resnik, 1998, 1999), publication timestamps (Papavassiliou et al., 2016), and structural or tag-related features within HTML pages (Chen and Nie, 2000; Resnik and Smith, 2003; Shi et al., 2006). These techniques were later addressed as a topic modeling task (Zafarian et al., 2015). While metadata often offers useful signals for identifying the documents, it alone cannot be considered to determine the alignment, disregarding the textual content in the document. Furthermore, metadata features do not always transfer reliably across different domains or web platforms.

An alternative line of research builds upon translation-based cues to determine the aligned documents. These primarily rely on quantifying the degree of translational equivalence between the source and target side documents. Some strategies incorporate bilingual dictionaries to identify cross-lingual lexical matches within the texts (Li and Gaussier, 2013). Word-level alignment has been examined by Ma and Liberman (1999); Fung and Cheung (2004); Ion et al. (2011), while phrase-level alignment has been explored by Gomes and Lopes (2016); Etchegoyhen and Gete (2020). Others have analyzed the top-ranked translation candidates within the document content itself (Morin et al., 2015; Espla-Gomis et al., 2016). A different approach involves translating non-English content into English and evaluating alignment through machine translation scoring metrics, as demonstrated by Uszkoreit et al. (2010); Zafarian et al. (2015); Rajitha et al. (2020). A limitation is that these techniques are sensitive to the quality of translation models or alignment algorithms. As such, their effectiveness can deteriorate for low-resource languages.

More recently, vector-based approaches have been introduced, where each document was represented as a vector, and semantic similarity was determined using a distance scoring function. Pairs that exceed a predefined similarity threshold were classified as aligned. Early techniques utilized models such as bag-of-words, TF-IDF (Jakubina and Langlais, 2016; Medved' et al., 2016; Buck and Koehn, 2016b; Germann, 2016), and word n-grams (Dara and Lin, 2016) to construct these document vectors.

Building on these efforts, El-Kishky and Guzmán (2020) employed multiPLMs to obtain vector representations for documents in multiple languages. They calculated inter-

document distances to identify translation pairs, using LASER2 embeddings (Artetxe and Schwenk, 2019b) to obtain representations. They have evaluated their strategy over a range of low-resource to high-resource languages. This technique forms the baseline for the current research, and a detailed explanation is provided in Section 4.3.3.1.

### 4.2.2 Sentence Alignment

Sentence alignment involves detecting sentence pairs across two languages that are either complete or partial translations of one another. Early work relied on statistical techniques such as sentence-length ratios (Brown et al., 1991; Gale and Church, 1993). However, the effectiveness of this method diminished significantly when there was a weak correlation between sentence lengths in the source and target languages (Ma, 2006). To address these shortcomings, later approaches integrated bilingual lexicons (Varga et al., 2007), probabilistic models for word or phrase alignment (Fung and Cheung, 2004; Etchegoyhen and Gete, 2020), and methods utilizing bilingual suffix trees for sentence and phrase pairing (Munteanu and Marcu, 2002). In another line of work, Stefanescu et al. (2012) framed the sentence alignment task as an information retrieval problem, while Munteanu and Marcu (2005) applied supervised classification techniques. Additional strategies involved translating the source sentences into English using a translation model and then applying retrieval-based methods to find their translation counterparts (Abdul-Rauf and Schwenk, 2009; Sarikaya et al., 2009; Mahata et al., 2017; Azpeitia et al., 2017, 2018).

Subsequent work has focused on using sentence or word embeddings to represent the source and target sentences in a shared semantic space. Once embedded representations are obtained, sentence similarity is computed to determine the best aligned pairs. Initial models for generating embeddings used techniques such as bi-directional Recurrent Neural Networks (RNNs) (Grégoire and Langlais, 2017), Deep Averaging Networks (DANs) (Iyyer et al., 2015), bi-gram driven neural architectures (Guoa et al., 2018), and auto-encoders (Leong et al., 2018). Some hybrid models further refined alignment decisions by applying supervised classifiers over the similarity scores produced by embedding-based methods (Bouamor and Sajjad, 2018; Leong et al., 2018).

Improving sentence alignment can be approached either by enhancing the underlying sentence representations or by refining the similarity scoring mechanisms. In the context of representation learning, Artetxe and Schwenk (2019b) employed LASER2 multilingual embeddings trained in a supervised manner, while (Kvapilíková et al., 2020) leveraged unsupervised embeddings to determine the sentence alignment. Although cosine similarity is a common unsupervised metric used to determine the proximity between sentence vectors, it often falls short in accuracy. Consequently, several improved semantic similarity metrics have been proposed (Guoa et al., 2018; Hangya and Fraser, 2019; Artetxe and Schwenk, 2019a). The approach introduced by Artetxe and

Schwenk (2019a), which we adopt as the baseline for our system, is examined further in Section 4.3.4.1.

### 4.2.3 Pre-trained Multilingual Language Models (multiPLMs)

As highlighted in the previous sections, sentence embeddings produced by multiPLMs have immensely contributed towards the success of both document alignment and sentence alignment tasks. LASER2 multiPLM, has been widely employed in large-scale bitext mining initiatives such as ParaCrawl (Bañón et al., 2020), wikiMatrix (Schwenk et al., 2021a), and CCMatrix (Schwenk et al., 2021b). Other multiPLMs such as mBERT (Devlin et al., 2019b), XLM-R (Conneau et al., 2020b) and LaBSE (Feng et al., 2022) have demonstrated strong performance on cross-lingual tasks.

### 4.2.4 Evaluating Document Alignment and Sentence Alignment Tasks

The evaluation of sentence and document alignment systems requires high-quality benchmark datasets. One such resource for evaluating document-level alignment methods is the corpus introduced by Buck and Koehn (2016a) offered for the WMT16 Bilingual Document Alignment Shared Task. It was a manually aligned English–French dataset annotated by humans. Rather than relying on fully human-annotated sentence alignments, Zweigenbaum et al. (2018) proposed a novel evaluation framework for the sentence alignment shared. They artificially injected known parallel sentences into the monolingual sides of comparable corpora to identify the aligned sentence pairs during the task. This enabled automatic evaluation, avoiding manual annotations. In several shared tasks, such as Koehn et al. (2018) and Koehn et al. (2019), manually aligned sentence-level gold standards were not provided. Instead, evaluation was carried out by measuring the impact of sentence alignment quality on downstream NMT systems.

For the evaluation of document-alignment and sentence-alignment tasks between `Sinhala-Tamil` and `English` languages, such a benchmark evaluation set is not available.

## 4.3 Methodology

This section covers dataset preparation (Section 4.3.1), as well as our algorithms to incorporate bilingual lexicons to improve document alignment (Section 4.3.3) and sentence alignment (Section 4.3.4).

### 4.3.1 Dataset

In this section, we describe the training datasets and the evaluation datasets which were used in this research work.

#### 4.3.1.1 Preparing Document and Sentence Alignment Evaluation Dataset

We expanded the dataset that was initially created by Rajitha et al. (2020), to make it a balanced dataset across the four news domains. The initial evaluation set was compiled by Rajitha et al. (2020) had a data imbalance from the NewsFirst and ITN news sources. Therefore for both document alignment and sentence alignment evaluation sets, we have crawled more data to make them contain data in similar quantities. However the Tamilnews updates were less frequent in the ITN website. Hence compared to the Army, Hiru and NewsFirst news sources, the ITN evaluation set contain less data for English-Tamil and Sinhala-Tamillanguage-pairs. For the purpose of expanding the dataset, we follow the steps described in this section. From the news sources, Hiru News[1], NewsFirst[2], Army News[3], and ITN[4], we web crawled news data from January 2013 through to April 2021.

We followed the initial pre-processing pipeline, the paragraph-level content from each webpage was concatenated into a continuous string. Non-textual data embedded within images and video tags was discarded. Additionally, documents containing fewer than fifty tokens were excluded to filter out extremely short content.

Among the selected websites, Army News, Hiru, and NewsFirst exhibit a high degree of consistency across languages in terms of structure, content scope, and sentence sequencing. As a result, direct translations of most English news articles were available in both Sinhala and Tamil. However, ITN displayed inconsistencies; equivalent articles in English were often missing in the other two languages, and as a result, we were limited with the ITN news data for the English-Sinhala and English-Tamil language-pairs.

For initial document alignment, site-specific heuristics tailored to each platform were applied. These candidate alignments were then reviewed and validated by human annotators fluent in the respective languages. The manually confirmed alignments were compiled into a gold standard evaluation set for use in benchmarking. The final statistics of the document evaluation dataset is shown in Table 4.1.

Table 4.1: Statistics of document alignment evaluation dataset

| Website | Sinhala - English | | | Tamil - English | | | Sinhala - Tamil | | |
|---|---|---|---|---|---|---|---|---|---|
| | Si | En | Aligned | Ta | En | Aligned | Si | Ta | Aligned |
| Army | 2,033 | 2,081 | 1,848 | 1,905 | 2,081 | 1,671 | 2,033 | 1,905 | 1,578 |
| Hiru | 3,133 | 1,634 | 1,397 | 2,886 | 1,634 | 1,056 | 3,133 | 2,886 | 2,002 |
| ITN | 6,641 | 3,212 | 1,150 | 3,035 | 3,212 | 707 | 6,641 | 3,035 | 979 |
| NewsiFrst | 3,936 | 4,273 | 1,680 | 3,929 | 3,228 | 1,266 | 3,936 | 3,929 | 1,433 |

[1] http://www.hirunews.lk
[2] https://www.newsfirst.lk/
[3] https://www.army.lk/
[4] https://www.itnnews.lk

The sentence alignment task was carried out using the document pairs confirmed in the gold-standard alignment set. Table 4.2 presents the total number of source and target sentences available for each language pair. Due to the extensive volume of sentences on both sides, a complete manual annotation of all possible translation pairs was impractical. To address this, a curated evaluation set comprising 300 one-to-one aligned sentence pairs per language pair was constructed from each selected website.

Table 4.2: Statistics of the sentence alignment evaluation dataset

| Language Pair | No. of Source Sentences | No. of Target Sentences |
|---|---|---|
| Sinhala-English | 153,750 | 140,701 |
| Tamil-English | 87,266 | 87,330 |
| Sinhala-Tamil | 38,101 | 37,371 |

#### 4.3.1.2 Parallel Data

As parallel data, we considered the bilingual lexicons: person names, designations, word dictionaries and glossaries. The `English-Sinhala` and `English-Tamil` Person Names and Designation lists were from the work of Priyadarshani et al. (2019), while the `Sinhala-Tamil` bilingual lists were from (Farhath et al., 2018a). The bilingual dictionaries have been extracted and used internally in an independent research and are yet to be published. A part of the `Tamil-English` dictionary is available at the WMT 2020 shared task[5]. We have obtained a Trilingual Glossary[6] from the Department of Official Languages, Sri Lanka. Statistics and samples of these bilingual lexicons are shown in Table 4.3 and Table 4.4, respectively.

Table 4.3: Statistics of the Bilingual Lexicons

| | No of Terms | | |
|---|---|---|---|
| Bilingual Lexicon | Sinhala-English | Tamil-English | Sinhala-Tamil |
| Person Names | 6,194 | 1,374 | 76,334 |
| Designations | 6,764 | 5,779 | 44,193 |
| Dictionary | 23,722 | 36,551 | 19,132 |
| Glossary | 24,261 | 24,261 | 24,261 |

#### 4.3.1.3 Dataset to Evaluate Downstream NMT Performance

We evaluate our NMT models against the SiTa-Trilingual evaluation set (Fernando et al., 2020) and Flores-v1 (Guzmán et al., 2019) benchmark evaluation sets, respec-

[5] `http://www.statmt.org/wmt20/translation-task.html`
[6] `https://www.languagesdept.gov.lk/`

Table 4.4: Overview of the Bilingual Lexicons

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

tively. Both these evaluation sets are gold-standard human-curated datasets which are commonly used for evaluating Sinhala, Tamil and English related NMT models.

### 4.3.2 Justification for Selecting MultiPLMs

We use encoder-based multiPLMs, LASER2, XLM-R and LaBSE to obtain embeddings for English-Sinhala, English-Tamil and Sinhala-Tamillanguage-pairs. Although mBERT (Devlin et al., 2019a) and XLM (Conneau and Lample, 2019) are alternatives, since they do not cover Sinhala in the case of LASER2 and both Sinhala and Tamil in the case of XLM, they were not considered to obtain embeddings.

**LASER2:** (L=2, H=1024, BiLSTM) is the shared encoder of a sequence-to-sequence BiLSTM NMT system trained on 93 languages. This model had been pre-trained only with parallel data. Due to training on parallel data, it has been favourable for cross-lingual tasks such as document alignment and sentence alignment tasks (El-Kishky et al., 2020; Bañón et al., 2020).

**XLM-R:** (L=12, H=768, A=6, 278M) is a multiPLM trained purely on a massive monolingual data collection of 2TB of filtered common crawl data using the Masked Language Model (MLM) objective. It supports 100 languages and have proved performance on cross-lingual Named Entity Recognition and question answering tasks (Conneau et al., 2020a).

**LaBSE:** (L=12, H=768, A=12, 471M) is pre-trained and fine-tuned multiPLM covering 104 languages. It had undergone pre-training on monolingual data and fine-tuned on 10 Million parallel data to oprimze for sentence-retrieval tasks such as bitext mining. With the fine-tuninig phase, the multiPLM embeddings are improved to support sentence-retrieval tasks such as bitext mining (Feng et al., 2020). The LaBSE model has been employed in several bitext mining tasks (Gala et al., 2023) even in LRL setting. Due to the coverage of the three language pairs under our study, and due to the differences in the pre-training stage, we select these three multiPLMs in our empirical study.

### 4.3.3 Document Alignment

In this section, we describe the document alignment algorithm proposed by El-Kishky and Guzmán (2020) and the improvement proposed by Rajitha et al. (2020), which were used to evaluate the performance of the multiPLMs.

#### 4.3.3.1 Baseline Implementation

We consider the work of El-Kishky and Guzmán (2020) to create the document alignment baseline system for our dataset. Their work introduces a (1) distance-scoring function to calculate the semantic similarity between the source and target documents and (2) a document matching algorithm, which produces the final aligned document pairs.

**Distance Scoring Function** El-Kishky and Guzmán (2020) introduced a novel distance metric named Cross-Lingual Sentence Mover's Distance (XLSMD), based on LASER2 multilingual embeddings to determine the semantic similarity between two documents. XLSMD is a distance metric based on Earth Mover's Distance (EMD). XLSMD represents each document as normalized bag-of-sentences (nBOS) with all the sentences containing a pre-calculated probability mass (weight). Equation 4.1 shows the semantic distance between documents $A$ and $B$. Here, $\Delta(i, j)$ is the Euclidean distance between the two sentence embeddings. As explained in Equation 4.2, $T_{i,j}$ is how much of sentence $i$ in document $A$ is assigned to sentence $j$ in document $B$ (probability mass of a sentence).

$$XLSMD(A, B) = \min_{T \geqslant 0} \sum_{i=1}^{V} \sum_{j=1}^{V} T_{i,j} \times \Delta(i, j) \tag{4.1}$$

$$Subject\ to : \forall i \sum_{j=1}^{V} T_{i,j} = d_{A,i} \ \ , \ \ \forall j \sum_{i=1}^{V} T_{i,j} = d_{B,j} \tag{4.2}$$

Equation 4.3 shows the first function used for the probability mass calculation. Here, they had used the relative frequencies of sentences as the probability mass. In this, $\sum_{s \in A} count(s)$ represents the sentence count in document $A$. After calculating XLSMD, the distance is used in the document matching algorithm.

$$d_{A,i} = \frac{count(i)}{\sum_{s \in A} count(s)} \tag{4.3}$$

To make the XLSMD calculations more tractable, the greedy algorithm introduced by El-Kishky and Guzmán (2020), Greedy Mover's Distance(GMD), an alternate to the relaxed-EMD was used. Here, the algorithm first calculates the Euclidean distance between each sentence pair and sorts them in ascending order. Then it iteratively multiplies each distance by the smallest weight among the two sentences, which is named as the flow value as shown in Equation 4.4.

$$distance = distance + ||s_A - s_B|| \times flow \times w_{A,B} \tag{4.4}$$

**Sentence Length (SL) Weighting**

This weighting scheme is used under the assumption that longer sentences should be given more probability mass than shorter sentences. Equation 4.5 defines how this weight is calculated.

$$d_{A,i} = \frac{count(i) \times |i|}{\sum_{s \in A} count(s) \times |s|} \tag{4.5}$$

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

**IDF Weighting**

IDF stands for Inverse Document Frequency. Here, they have used the argument that the sentences that occur more frequently in the corpus should be given less importance than the infrequent sentences in the document. Equation 4.6 defines how it is calculated.

$$d_{A,i} = 1 + \log \frac{N + 1}{1 + |d \in D : s \in d|} \tag{4.6}$$

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

**SLIDF Weighting**

In this scheme, both the above schemes are joined together to form a joint weighting scheme. It is shown in Equation 4.7.

$$d_{A,i} = SL(i) * IDF(i) \tag{4.7}$$

**Document matching Algorithm** We follow the same document matching algorithm (El-Kishky and Guzmán, 2020) used a document matching algorithm to obtain the final aligned document pairs. In this algorithm, initially, the semantic distances between each source document and target documents are calculated according to the above-mentioned scoring function. Then, starting from the document pair containing the minimum distance, subsequent pairs $d_A$ and $d_B$ are selected iteratively, such that the documents $d_A$ and $d_B$ have not been considered in a previous selection.

#### 4.3.3.2 Weighting Scheme based on Bilingual Lexicons

Rajitha et al. (2020) modified the distance scoring function of El-Kishky and Guzmán (2020), by introducing a new weighting scheme considering bilingual lexicons. This weight calculation differs based on the nature of the term mapping in the bilingual lexicons (as word-to-word mappings or phrase-to-phrase mappings). This is described in the following section. With this improvement, the semantic distance calculation between a source side document $d_A$ and target side document $d_B$ is shown in Figure 4.1.

We use the bilingual lexicons mentioned in Section 4.3.1 to introduce a weighting

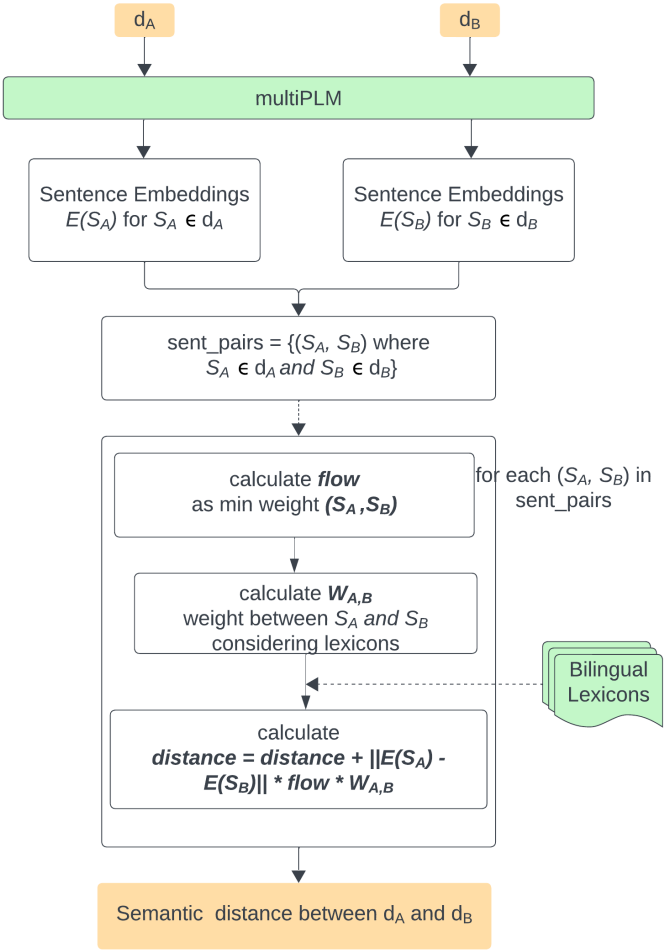


Figure 4.1: Process for calculating the semantic distance between source language document $d_A$ and target language document $d_B$. Here $w_{A,B}$ refers to the weight considering bilingual lexicons between sentence $s_A$ and $s_B$. The semantic distance scored from this process would be used by the Document matching algorithm (Section 4.3.3.1) to finally produce the aligned document pairs.

scheme on top of the SL, IDF and SLIDF schemes. Here, if a sentence $s_A$ from document $A$ contains a word $w$ in the bilingual lexicon and the sentence $s_B$ from document $B$ contains the translation of the word $w$, a counter value is incremented. The calculation considering this lexicon term count is shown in Equation 4.8.

$$w_{A,B} = \frac{|s_A| - count}{|s_A|} \qquad |s_A| = Number\ of\ tokens\ in\ sentence\ s_A \tag{4.8}$$

$w_{A,B}$ is the weighting introduced between sentence $s_A$ and $s_B$, and this is included into the Greedy Mover's Distance (GMD) algorithm as shown in Equation 4.9 to finally calculate the semantic distance between the sentences.

$$distance = distance + \|s_A - s_B\| \times flow \times w_{A,B} \tag{4.9}$$

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

**Usage of Person Names Bilingual List** Similar to Rajitha et al. (2020), we added the parallel words in person names bilingual list into a dictionary data structure where keys are words from language $A$ and the values are arrays of translations of the key in language $B$ (One person name has multiple translations sometimes due to multiple types of spelling formats). When calculating the weights, for each sentence pair, we iterated through the words in the sentence to calculate the mapping counts. Here, we split the sentence $s_A$ into words and check if each word $w$ exists in the dictionary. If it exists, we get the parallel words $v_B$, and check if each parallel word exists in the sentence $s_B$. If so, we increase the counter and remove the mapped word from the sentence $s_B$. This counter value is used as the input in Equation 4.8. Algorithm 1 explains this process.

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

```
Require: s_A, s_B, dict
w_A ⇐ list(s_A)
w_B ⇐ list(s_B)
count ⇐ 0
for w ∈ w_A : |w| = 1 do
    if w ∈ dict then
        v_B ⇐ dict[w]
        for v ∈ v_B do
            if v ∈ w_B then
                count ⇐ count + 1
                Remove w from w_B
            end if
        end for
    end if
end for
```

**Usage of Designations Bilingual List and Word Dictionary** Different to the person names, bilingual lists and word dictionaries contain phrases. Therefore, when calculating weights, we implemented a separate algorithm that identifies the multiple word mappings considering the multiple words. Similar to Rajitha et al. (2020), for each sentence $s_A$, we get all the permutations of words from length one to length five (the maximum length of a record in the dictionary is five). Then we do the same process described above to get the mapping counts. Algorithm 2 depicts this process. When person names, designations, and word dictionaries are used in combination, we sum up the counter values from both Algorithm 1 and 2, use that value as the input for Equation 4.8.

**Improved Dictionary** We use the improved dictionary produced by (Rajitha et al., 2020) to be used as the final parallel data to consider with the improved scoring function. An overview of the improved dictionary is shown in Table 4.5.

**Algorithm 2:** Calculate count for Equation 8 considering Person Names lexicon

```
Require: s_A, s_B, dict
w_A ⇐ list(s_A)
w_B ⇐ list(s_B)
count ⇐ 0
if |w_A| ⩾ 5 then
    for w ∈ w_A : |w| = 1, 2, 3, 4, 5 do
        if w ∈ dict then
            v_B ⇐ dict[w]
            for v ∈ v_B do
                if v ∈ w_B then
                    count ⇐ count + 1
                    Remove w from w_B
                end if
            end for
        end if
    end for
else
    Algorithm 1
end if
```

1

Table 4.5: Overview of the Improved Dictionary

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

### 4.3.4 Sentence Alignment

In this section, we describe the sentence alignment algorithm proposed by Artetxe and Schwenk (2019a) and the improvement proposed by Rajitha et al. (2020), which were used to evaluate the impact of the selected multiPLMs.

#### 4.3.4.1 Baseline system

Our baseline is the system proposed by Artetxe and Schwenk (2019a). They obtained the LASER2 multilingual sentence embeddings for all the source and target sentences, and aligned these sentence embeddings considering a margin-based cosine similarity. This similarity measurement considers a margin between the cosine of a given sentence pair and that of its respective nearest neighbours. They proposed the following three criteria for determining the candidate translation sentence, focusing on a higher recall at the cost of precision.

- **Forward**: For each source side sentence, the highest scoring sentence is selected from the target sentences, as its target side translation.
- **Backwards**: For each target-side sentence, the highest-scoring source sentence is selected as its source-side translation sentence.

- **Intersection**: Intersection of the sentence pairs identified from the forward and backwards criteria.

#### 4.3.4.2 Sentence Similarity Scoring

Here we use the improved semantic distance measurement proposed by Rajitha et al. (2020). They introduced a weighing, on top of the calculated sentence similarity using bilingual lexicons (Section 4.3.1). Similar to document alignment, we consider the bilingual lists, person names, designations, dictionary and improved dictionary as parallel data with this improved weighting scheme. When sentences in the source language document $d_A$ and sentences in the target language document $d_B$ are given as inputs, the sentence alignment algorithm would produce the aligned parallel sentence pairs, as shown in Figure 4.2.

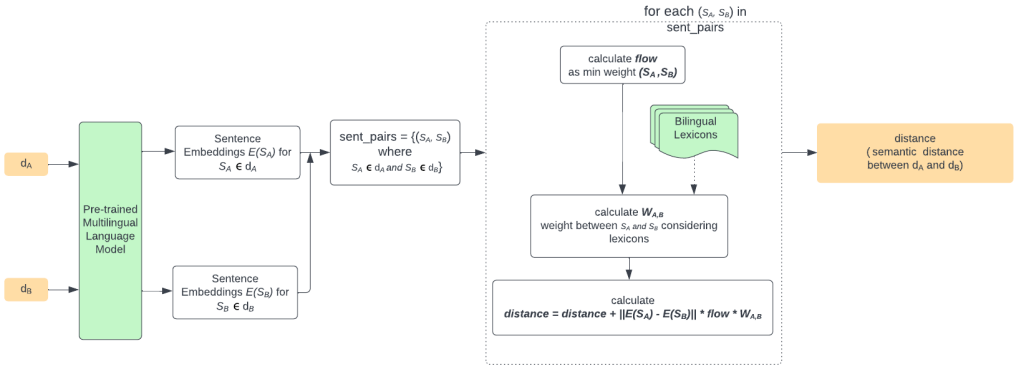


Figure 4.2: Given the source and target language sentences, the diagram outlines the sentence alignment algorithm considering the forward criterion. In the backwards criterion, for each $s_B$ in $d_B$, the aligned sentences are picked up from the source side.

In the forward criterion, we use the cosine similarity as the initial similarity score and select the best matching neighbourhood (k) as 4 candidates, similar to Artetxe and Schwenk (2019a) [7], for each source sentence. Here, if the source sentence $s_A$ from document $A$ contains a word $w$ in the parallel dataset and the target sentence $s_B$ from the selected k candidates contains the translation of the word $w$, a counter value is incremented. This counter value is used to calculate the weight using Equation 4.10 (Multiplicative inverse of Equation 4.8), which gives a higher weight for sentence pairs having more overlapping tokens and a lower weight for sentence pairs with a lower number of overlapping tokens.

$$w_{A,B} = \frac{|s_A|}{|s_A| - count} \qquad |s_A| = Number\ of\ tokens\ in\ source\ sentence\ s_A \quad (4.10)$$

[7] Artetxe and Schwenk (2019a) have experimented with different k values and have selected 4 as the preferred value

New similarity score between each source sentence $s_A$ and each target sentence $s_B$, according to the selected $k$ candidates is calculated using Equation 4.11.

$$similarity\_score_{A,B} = cosine\_similarity_{A,B} \times w_{A,B} \tag{4.11}$$

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

**Date-wise Filtering** Since our dataset was completely taken from the news domain, all the news documents have the published date as metadata. Moreover, in most cases, the same news document is published in all three languages on the same day. Therefore, before starting the alignment process, we filtered and divided the documents using the published date and reduced the search space by a considerable amount.

## 4.4 Experiments and Results

We evaluate the effect of the selected multiPLMs separately for document alignment and sentence alignment tasks, using the extended golden alignment dataset we prepared (see Section 4.3.1). Further, an extrinsic evaluation was conducted on the sentence alignment task by training NMT models (Section 4.4.3).

### 4.4.1 Document Alignment

For document alignment, we report the results for the baseline system using the technique proposed by El-Kishky and Guzmán (2020) with LASER2 multilingual embeddings for each news web source. Subsequently, an ablation study was conducted by sequentially adding each bilingual lexicon on top of the previous experiment. Then we repeat the above experiments for XLM-R and LaBSE. Thus, this becomes the first empirical study of these three models for the task of document alignment.

Our technique is aimed at high recall at the cost of low precision. However, we have reported the Recall (R), Precision (P) and F1 scores over the gold-standard evaluation set. We experimented for `English-Sinhala`, `English-Tamil` and `Sinhala-Tamil` language pairs for each news web source. The results are shown in Table 4.6 for English-Sinhala, Table 4.7 for `English-Tamil` and Table 4.8 for `Sinhala-Tamil`. In the tables, experiments with Person Names lists *(BL+P)*, Person Names with Designation lists *(BL+P+Ds)*, Person Names, Designation lists with Dictionary *(BL+P +Ds+Dc)* and Person Names, Designation lists, Dictionary with improved dictionary *(BL+P+Ds+MDc)* were repeated considering embeddings from LASER2, XLM-R and LaBSE.

From the results, it was observed that the scores obtained for the baseline were mostly outperformed by incorporating the bilingual lists, irrespective of the type of embeddings used for the `English-Sinhala` language pair. The highest gain was for

Table 4.6: Document Alignment results in terms of Precision(P), Recall (R) and F1 for English-Sinhala language pair.

| | | | En-Si | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Hiru | | | ITN | | | Newsfirst | | | Army | | |
| | | | R | P | F1 | R | P | F1 | R | P | F1 | R | P | F1 |
| LASER | | | | | | | | | | | | | | |
| A | Baseline | SL | 82.25 | 71.06 | 76.24 | 91.22 | 37.28 | 52.93 | 96.01 | 47.37 | 63.44 | 99.35 | 94.49 | 96.86 |
| | | IDF | 79.31 | 68.52 | 73.52 | 89.39 | 36.53 | 51.87 | 94.17 | 46.46 | 62.22 | 97.02 | 92.28 | 94.59 |
| | | SLIDF | 82.32 | 71.12 | 76.31 | 91.22 | 37.28 | 52.93 | 95.89 | 47.31 | 63.36 | 99.35 | 94.49 | 96.86 |
| B | Names | SL | 84.90 | 73.35 | 78.70 | 92.78 | 37.92 | 53.84 | 96.31 | 47.52 | 63.64 | 99.19 | 94.34 | 96.70 |
| | | IDF | 81.89 | 70.75 | 75.91 | 90.78 | 37.10 | 52.67 | 94.17 | 46.46 | 62.22 | 97.73 | 92.95 | 95.28 |
| | | SLIDF | 84.90 | 73.35 | 78.70 | 92.87 | 37.95 | 53.88 | 96.31 | 47.52 | 63.64 | 99.19 | 94.34 | 96.70 |
| C | Names+Desig | SL | 84.90 | 73.35 | 78.70 | 92.78 | 37.92 | 53.84 | 96.31 | 47.52 | 63.64 | 99.19 | 94.34 | 96.70 |
| | | IDF | 81.89 | 70.75 | 75.91 | 90.78 | 37.10 | 52.67 | 94.17 | 46.46 | 62.22 | 97.73 | 92.95 | 95.28 |
| | | SLIDF | 84.90 | 73.35 | 78.70 | 92.87 | 37.95 | 53.88 | 96.31 | 47.52 | 63.64 | 99.19 | 94.34 | 96.70 |
| D | Names+Desig+Dic | SL | 85.61 | 73.96 | 79.36 | 93.13 | 38.06 | 54.04 | 96.55 | 47.64 | 63.80 | **99.41** | **94.55** | **96.91** |
| | | IDF | 81.89 | 70.75 | 75.91 | 90.78 | 37.10 | 52.67 | 94.17 | 46.46 | 62.22 | 97.73 | 92.95 | 95.28 |
| | | SLIDF | 84.90 | 73.35 | 78.70 | 92.87 | 37.95 | 53.88 | 96.31 | 47.52 | 63.64 | 99.19 | 94.34 | 96.70 |
| E | Names+Desig+modDic | SL | **85.90** | **74.21** | **79.63** | **94.00** | **38.41** | **54.54** | **97.32** | **48.02** | **64.31** | **99.41** | **94.55** | **96.91** |
| | | IDF | 81.89 | 70.75 | 75.91 | 90.78 | 37.10 | 52.67 | 94.17 | 46.46 | 62.22 | 97.73 | 92.95 | 95.28 |
| | | SLIDF | 84.90 | 73.35 | 78.70 | 92.87 | 37.95 | 53.88 | 96.31 | 47.52 | 63.64 | 99.19 | 94.34 | 96.70 |
| XLMR | | | | | | | | | | | | | | |
| A | Baseline | SL | 91.05 | 78.66 | 84.41 | 98.09 | 40.09 | 56.91 | 98.39 | 48.55 | 65.01 | 99.46 | 94.60 | 96.97 |
| | | IDF | 91.41 | 78.97 | 84.74 | 98.00 | 40.05 | 56.86 | 98.21 | 48.46 | 64.90 | 99.03 | 94.18 | 96.54 |
| | | SLIDF | 90.91 | 78.54 | 84.27 | 98.09 | 40.09 | 56.91 | 98.39 | 48.55 | 65.01 | 99.46 | 94.60 | 96.97 |
| B | Names | SL | 92.77 | 80.15 | 86.00 | **98.26** | **40.16** | **57.02** | 98.57 | 48.63 | 65.13 | **99.73** | **94.85** | **97.23** |
| | | IDF | 92.27 | 79.72 | 85.54 | 97.83 | 39.98 | 56.76 | 97.92 | 48.31 | 64.70 | 99.35 | 94.49 | 96.86 |
| | | SLIDF | 92.91 | 80.27 | 86.13 | **98.26** | **40.16** | **57.02** | 98.57 | 48.63 | 65.13 | **99.73** | **94.85** | **97.23** |
| C | Names+Desig | SL | 92.77 | 80.15 | 86.00 | **98.26** | **40.16** | **57.02** | 98.57 | 48.63 | 65.13 | **99.73** | **94.85** | **97.23** |
| | | IDF | 92.27 | 79.72 | 85.54 | 97.83 | 39.98 | 56.76 | 97.92 | 48.31 | 64.70 | 99.35 | 94.49 | 96.86 |
| | | SLIDF | 92.91 | 80.27 | 86.13 | **98.26** | **40.16** | **57.02** | 98.57 | 48.63 | 65.13 | **99.73** | **94.85** | **97.23** |
| D | Names+Desig+Dic | SL | 92.63 | 80.03 | 85.87 | 98.26 | 40.16 | 57.01 | 98.51 | 48.60 | 65.09 | **99.73** | **94.85** | **97.23** |
| | | IDF | 92.27 | 79.72 | 85.54 | 97.83 | 39.98 | 56.76 | 97.92 | 48.31 | 64.70 | 99.35 | 94.49 | 96.86 |
| | | SLIDF | 92.91 | 80.27 | 86.13 | **98.26** | **40.16** | **57.02** | 98.57 | 48.63 | 65.13 | **99.73** | **94.85** | **97.23** |
| E | Names+Desig+modDic | SL | **93.63** | **80.89** | **86.80** | 98.17 | 40.12 | 56.96 | **98.69** | **48.69** | **65.21** | **99.73** | **94.85** | **97.23** |
| | | IDF | 92.27 | 79.72 | 85.54 | 97.83 | 39.98 | 56.76 | 97.92 | 48.31 | 64.70 | 99.35 | 94.49 | 96.86 |
| | | SLIDF | 92.91 | 80.27 | 86.13 | **98.26** | **40.16** | **57.02** | 98.57 | 48.63 | 65.13 | **99.73** | **94.85** | **97.23** |
| LABSE | | | | | | | | | | | | | | |
| A | Baseline | SL | 95.42 | 82.44 | 88.45 | 98.78 | 40.37 | 57.32 | **99.11** | **48.90** | **65.49** | **99.73** | **94.85** | **97.23** |
| | | IDF | 95.49 | 82.50 | 88.52 | 98.35 | 40.19 | 57.06 | 99.23 | 48.96 | 65.56 | 99.67 | 94.80 | 97.18 |
| | | SLIDF | 95.35 | 82.38 | 88.39 | 98.78 | 40.37 | 57.32 | **99.11** | **48.90** | **65.49** | **99.73** | **94.85** | **97.23** |
| B | Names | SL | 95.42 | 82.44 | 88.46 | 98.87 | 40.41 | 57.37 | 98.99 | 48.84 | 65.41 | **99.73** | **94.85** | **97.23** |
| | | IDF | **95.71** | **82.68** | **88.72** | 98.43 | 40.23 | 57.12 | 98.99 | 48.84 | 65.41 | 99.68 | 94.80 | 97.18 |
| | | SLIDF | 95.42 | 82.44 | 88.46 | 98.87 | 40.41 | 57.37 | 98.99 | 48.84 | 65.41 | **99.73** | **94.85** | **97.23** |
| C | Names+Desig | SL | 95.42 | 82.44 | 88.46 | 98.87 | 40.41 | 57.37 | 98.99 | 48.84 | 65.41 | **99.73** | **94.85** | **97.23** |
| | | IDF | **95.71** | **82.68** | **88.72** | 98.43 | 40.23 | 57.12 | 98.99 | 48.84 | 65.41 | 99.68 | 94.80 | 97.18 |
| | | SLIDF | 95.42 | 82.44 | 88.46 | 98.87 | 40.41 | 57.37 | 98.99 | 48.84 | 65.41 | **99.73** | **94.85** | **97.23** |
| D | Names+Desig+Dic | SL | 95.28 | 82.31 | 88.32 | 98.96 | 40.44 | 57.42 | **99.11** | **48.90** | **65.49** | **99.73** | **94.85** | **97.23** |
| | | IDF | **95.71** | **82.68** | **88.72** | 98.43 | 40.23 | 57.12 | 98.99 | 48.84 | 65.41 | 99.68 | 94.80 | 97.18 |
| | | SLIDF | 95.42 | 82.44 | 88.46 | 98.87 | 40.41 | 57.37 | 98.99 | 48.84 | 65.41 | **99.73** | **94.85** | **97.23** |
| E | Names+Desig+modDic | SL | 95.49 | 82.50 | 88.52 | **99.04** | **40.48** | **57.47** | 98.99 | 48.84 | 65.41 | **99.73** | **94.85** | **97.23** |
| | | IDF | **95.71** | **82.68** | **88.72** | 98.43 | 40.23 | 57.12 | 98.99 | 48.84 | 65.41 | 99.68 | 94.80 | 97.18 |
| | | SLIDF | 95.42 | 82.44 | 88.46 | 98.87 | 40.41 | 57.37 | 98.99 | 48.84 | 65.41 | **99.73** | **94.85** | **97.23** |

Table 4.7: Document Alignment results in terms of Precision(P), Recall (R) and F1 for English-Tamil language pair.

| | | | Hiru | | | ITN | | | Newsfirst | | | Army | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | **R** | **P** | **F1** | **R** | **P** | **F1** | **R** | **P** | **F1** | **R** | **P** | **F1** |
| **LASER** | | | | | | | | | | | | | | |
| A | Baseline | SL | 25.13 | 18.09 | 21.04 | 50.78 | 19.00 | 27.65 | 53.71 | 21.47 | 30.68 | 72.27 | 67.43 | 69.77 |
| | | IDF | 22.65 | 16.31 | 18.96 | 52.62 | 19.68 | 28.65 | 52.45 | 20.97 | 29.96 | 64.51 | 60.20 | 62.28 |
| | | SLIDF | 25.30 | 18.21 | 21.18 | 50.92 | 19.05 | 27.72 | 53.95 | 21.57 | 30.81 | 72.39 | 67.54 | 69.88 |
| B | Names | SL | 26.07 | 18.77 | 21.83 | 52.05 | 19.47 | 28.34 | 54.34 | 21.72 | 31.04 | 73.85 | 68.91 | 71.29 |
| | | IDF | 24.70 | 17.78 | 20.68 | 54.03 | 20.21 | 29.42 | 52.76 | 21.09 | 30.13 | 64.81 | 60.47 | 62.56 |
| | | SLIDF | 26.24 | 18.89 | 21.97 | 52.05 | 19.47 | 28.34 | 54.66 | 21.85 | 31.22 | 73.91 | 68.97 | 71.35 |
| C | Names+Desig | SL | 26.07 | 18.77 | 21.83 | 52.05 | 19.47 | 28.34 | 54.34 | 21.72 | 31.04 | 73.85 | 68.91 | 71.29 |
| | | IDF | 24.70 | 17.78 | 20.68 | 54.03 | 20.21 | 29.42 | 52.76 | 21.09 | 30.13 | 64.81 | 60.47 | 62.56 |
| | | SLIDF | 26.24 | 18.89 | 21.97 | 52.05 | 19.47 | 28.34 | 54.66 | 21.85 | 31.22 | 73.91 | 68.97 | 71.35 |
| D | Names+Desig+Dic | SL | 47.44 | 34.15 | 39.71 | 74.82 | 27.99 | 40.74 | 76.14 | 30.44 | 43.49 | 84.55 | 78.89 | 81.62 |
| | | IDF | 44.36 | 31.94 | 37.14 | 74.26 | 27.78 | 40.43 | 72.20 | 28.86 | 41.24 | 77.97 | 72.75 | 75.27 |
| | | SLIDF | 47.35 | 34.09 | 39.64 | 74.82 | 27.99 | 40.74 | 75.83 | 30.31 | 43.31 | 84.55 | 78.89 | 81.62 |
| E | Names+Desig+modDic | SL | 50.85 | 36.62 | 42.58 | **77.23** | **28.89** | **42.05** | 80.25 | 32.08 | 45.84 | **87.02** | **81.19** | **84.00** |
| | | IDF | 47.44 | 34.15 | 39.71 | 76.52 | 28.62 | 41.66 | 75.99 | 30.38 | 43.40 | 79.79 | 74.45 | 77.03 |
| | | SLIDF | **50.94** | **36.68** | **42.65** | **77.23** | **28.89** | **42.05** | **80.57** | **32.21** | **46.02** | **87.02** | **81.19** | **84.00** |
| **XLM-R** | | | | | | | | | | | | | | |
| A | Baseline | SL | 82.31 | 59.26 | 68.91 | 94.34 | 35.29 | 51.37 | 97.08 | 38.81 | 55.45 | 94.77 | 88.43 | 91.49 |
| | | IDF | 81.62 | 58.77 | 68.34 | 95.33 | 35.66 | 51.91 | 96.92 | 38.74 | 55.36 | 95.36 | 88.98 | 92.06 |
| | | SLIDF | 82.39 | 59.32 | 68.98 | 94.34 | 35.29 | 51.37 | 96.92 | 38.74 | 55.36 | 94.77 | 88.43 | 91.49 |
| B | Names | SL | 82.82 | 59.63 | 69.34 | 94.63 | 35.40 | 51.53 | 97.08 | 38.81 | 55.45 | 95.53 | 89.14 | 92.22 |
| | | IDF | 82.65 | 59.51 | 69.20 | 94.34 | 35.29 | 51.37 | 95.34 | 38.11 | 54.45 | 95.12 | 88.76 | 91.83 |
| | | SLIDF | 82.91 | 59.69 | 69.41 | 94.63 | 35.40 | 51.53 | 96.92 | 38.74 | 55.35 | 95.53 | 89.14 | 92.22 |
| C | Names+Desig | SL | 82.82 | 59.63 | 69.34 | 94.63 | 35.40 | 51.53 | 97.08 | 38.81 | 55.45 | 95.53 | 89.14 | 92.22 |
| | | IDF | 82.65 | 59.51 | 69.20 | 94.34 | 35.29 | 51.37 | 95.34 | 38.11 | 54.45 | 95.12 | 88.76 | 91.83 |
| | | SLIDF | 82.91 | 59.69 | 69.41 | 94.63 | 35.40 | 51.53 | 96.92 | 38.74 | 55.35 | 95.53 | 89.14 | 92.22 |
| D | Names+Desig+Dic | SL | 85.04 | 61.23 | 71.20 | 97.31 | 36.40 | 52.98 | 97.71 | 39.06 | 55.81 | 97.42 | 90.90 | 94.04 |
| | | IDF | 84.10 | 60.55 | 70.41 | 96.46 | 36.09 | 52.52 | 96.60 | 38.62 | 55.18 | 96.30 | 89.86 | 92.97 |
| | | SLIDF | 85.04 | 61.23 | 71.20 | 97.31 | 36.40 | 52.98 | **97.79** | **39.09** | **55.85** | 97.42 | 90.90 | 94.04 |
| E | Names+Desig+modDic | SL | **85.21** | **61.35** | **71.34** | **97.45** | **36.45** | **53.06** | 97.71 | 39.06 | 55.81 | **97.53** | **91.01** | **94.16** |
| | | IDF | 83.93 | 60.43 | 70.27 | 96.89 | 36.24 | 52.75 | 96.68 | 38.65 | 55.22 | 96.18 | 89.75 | 92.85 |
| | | SLIDF | **85.21** | **61.35** | **71.34** | **97.45** | **36.45** | **53.06** | 97.45 | 36.45 | 53.06 | **97.53** | **91.01** | **94.16** |
| **LaBSE** | | | | | | | | | | | | | | |
| A | Baseline | SL | **87.09** | **62.71** | **72.92** | 99.58 | 37.25 | 54.22 | 98.10 | 39.22 | 56.03 | 98.47 | 91.89 | 95.07 |
| | | IDF | 85.64 | 61.66 | 71.70 | 99.58 | 37.25 | 54.22 | **98.10** | **39.22** | **56.03** | 98.30 | 91.72 | 94.89 |
| | | SLIDF | 87.01 | 62.65 | 72.84 | 98.10 | **39.22** | **56.03** | **98.10** | **39.22** | **56.03** | **98.47** | 91.89 | 95.07 |
| B | Names | SL | 86.75 | 62.46 | 72.63 | 99.58 | 37.25 | 54.22 | 97.95 | 39.15 | 55.94 | 98.41 | 91.83 | 95.01 |
| | | IDF | 85.81 | 61.78 | 71.84 | 99.15 | 37.09 | 53.99 | 96.68 | 38.65 | 55.22 | 98.18 | 91.61 | 94.78 |
| | | SLIDF | 86.67 | 62.40 | 72.56 | 99.58 | 37.25 | 54.22 | 97.95 | 39.15 | 55.94 | 98.41 | 91.83 | 95.01 |
| C | Names+Desig | SL | 86.75 | 62.46 | 72.63 | 99.58 | 37.25 | 54.22 | 97.95 | 39.15 | 55.94 | 98.41 | 91.83 | 95.01 |
| | | IDF | 85.81 | 61.78 | 71.84 | 99.15 | 37.09 | 53.99 | 96.68 | 38.65 | 55.22 | 98.18 | 91.61 | 94.78 |
| | | SLIDF | 86.67 | 62.40 | 72.56 | 99.58 | 37.25 | 54.22 | 97.95 | 39.15 | 55.94 | 98.41 | 91.83 | 95.01 |
| D | Names+Desig+Dic | SL | 85.64 | 61.66 | 71.70 | 99.43 | 37.20 | 54.14 | 96.76 | 38.68 | 55.27 | 98.12 | 91.56 | 94.72 |
| | | IDF | 86.41 | 62.22 | 72.34 | **99.86** | 37.35 | 54.37 | 97.95 | 39.15 | 55.94 | 98.41 | 91.83 | 95.01 |
| | | SLIDF | 86.41 | 62.22 | 72.34 | **99.86** | 37.35 | 54.37 | 97.87 | 39.12 | 55.90 | 98.41 | **95.01** | **96.68** |
| E | Names+Desig+modDic | SL | 86.41 | 62.22 | 72.34 | 99.86 | 37.35 | 54.37 | 97.87 | 39.12 | 55.90 | 98.41 | **95.01** | **96.68** |
| | | IDF | 85.90 | 61.85 | 71.91 | 99.43 | 37.20 | 54.14 | 96.92 | 38.74 | 55.36 | 98.06 | 91.50 | 94.67 |
| | | SLIDF | 86.15 | 62.03 | 72.13 | **99.86** | 37.35 | 54.37 | 97.87 | 39.12 | 55.90 | 98.41 | 91.83 | 95.01 |

Table 4.8: Document Alignment results in terms of Precision (P), Recall (R) and F1 for Sinhala-Tamil language pair.

| | | | Hiru | | | ITN | | | Newsfirst | | | Army | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | **R** | **P** | **F1** | **R** | **P** | **F1** | **R** | **P** | **F1** | **R** | **P** | **F1** |
| **LASER** | | | | | | | | | | | | | | |
| A | Baseline | SL | 43.71 | 32.61 | 37.35 | 84.68 | 30.12 | 44.44 | 82.82 | 28.24 | 42.12 | 82.45 | 72.24 | 77.01 |
| | | IDF | 41.81 | 31.20 | 35.73 | 85.80 | 30.52 | 45.03 | 79.76 | 27.20 | 40.56 | 76.81 | 67.30 | 71.74 |
| | | SLIDF | 43.81 | 32.69 | 37.44 | 84.68 | 30.12 | 44.44 | 82.03 | 27.97 | 41.72 | 82.45 | 72.24 | 77.01 |
| B | Names | SL | 49.10 | 36.64 | 41.96 | 88.87 | 31.61 | 46.63 | 85.09 | 29.01 | 43.27 | 86.57 | 75.85 | 80.86 |
| | | IDF | 46.40 | 34.63 | 39.66 | 90.40 | 32.16 | 47.44 | 83.42 | 28.44 | 42.42 | 80.86 | 70.85 | 75.52 |
| | | SLIDF | 49.25 | 36.75 | 42.09 | 88.87 | 31.61 | 46.63 | 85.00 | 28.98 | 43.22 | 86.69 | 75.96 | 80.97 |
| C | Names+Desig | SL | 49.10 | 36.64 | 41.96 | 88.87 | 31.61 | 46.63 | 85.09 | 29.01 | 43.27 | 86.57 | 75.85 | 80.86 |
| | | IDF | 41.81 | 31.20 | 35.73 | 85.80 | 30.52 | 45.03 | 85.09 | 29.01 | 43.27 | 76.81 | 67.30 | 71.74 |
| | | SLIDF | 49.25 | 36.75 | 42.09 | 88.87 | 31.61 | 46.63 | 85.00 | 28.98 | 43.22 | 86.69 | 75.96 | 80.97 |
| D | Names+Desig+Dic | SL | 52.60 | 39.25 | 44.95 | 91.52 | 32.56 | 48.03 | 87.27 | 29.75 | 44.38 | 87.07 | 76.29 | 81.33 |
| | | IDF | 50.35 | 37.57 | 43.03 | 92.44 | 32.88 | 48.51 | 85.19 | 29.05 | 43.32 | 82.13 | 71.96 | 76.71 |
| | | SLIDF | 52.45 | 39.13 | 44.82 | 91.52 | 32.56 | 48.03 | 87.27 | 29.75 | 44.38 | 87.20 | 76.40 | 81.44 |
| E | Names+Desig+modDic | SL | **57.34** | **42.79** | **49.01** | 93.97 | 33.43 | 49.32 | **90.92** | **31.00** | **46.24** | **89.73** | **78.62** | **83.81** |
| | | IDF | 54.50 | 40.66 | 46.57 | **94.38** | **33.58** | **49.53** | 88.55 | 30.19 | 45.03 | 84.60 | 74.13 | 79.02 |
| | | SLIDF | 57.24 | 42.71 | 48.92 | 93.97 | 33.43 | 49.32 | **90.92** | **31.00** | **46.24** | 89.67 | 78.57 | 83.75 |
| **XLM-R** | | | | | | | | | | | | | | |
| A | Baseline | SL | 78.77 | 58.78 | 67.32 | 98.47 | 35.03 | 51.68 | 98.81 | 33.69 | 50.25 | 92.65 | 81.18 | 86.53 |
| | | IDF | 77.07 | 57.51 | 65.87 | 99.18 | 35.28 | 52.05 | 98.32 | 33.52 | 50.00 | 89.61 | 78.51 | 83.69 |
| | | SLIDF | 78.82 | 58.81 | 67.36 | 98.47 | 35.03 | 51.68 | 98.81 | 33.69 | 50.25 | 92.65 | 81.18 | 86.53 |
| B | Names | SL | 79.87 | 59.60 | 68.26 | 99.08 | 35.25 | 52.00 | 98.82 | 33.69 | 50.25 | 94.36 | 82.68 | 88.13 |
| | | IDF | 78.87 | 58.85 | 67.40 | 99.18 | 35.28 | 52.05 | 98.32 | 33.52 | 50.00 | 91.63 | 80.29 | 85.59 |
| | | SLIDF | 79.82 | 59.56 | 68.22 | 99.08 | 35.25 | 52.00 | 98.82 | 33.69 | 50.25 | 94.36 | 82.68 | 88.13 |
| C | Names+Desig | SL | 79.87 | 59.60 | 68.26 | 99.08 | 35.25 | 52.00 | 98.82 | 33.69 | 50.25 | 94.36 | 82.68 | 88.13 |
| | | IDF | 77.07 | 57.51 | 65.87 | 99.18 | 35.28 | 52.05 | 98.82 | 33.69 | 50.25 | 89.61 | 78.51 | 83.69 |
| | | SLIDF | 79.82 | 59.56 | 68.22 | 99.08 | 35.25 | 52.00 | 98.82 | 33.69 | 50.25 | 94.36 | 82.68 | 88.13 |
| D | Names+Desig+Dic | SL | 80.27 | 59.90 | 68.60 | 99.49 | 35.39 | 52.21 | 98.91 | 33.73 | 50.30 | 94.55 | 82.84 | 88.31 |
| | | IDF | 78.82 | 58.81 | 67.36 | 99.18 | 35.28 | 52.05 | 98.42 | 33.56 | 50.05 | 92.08 | 80.68 | 86.00 |
| | | SLIDF | 80.27 | 59.90 | 68.60 | 99.49 | 35.39 | 52.21 | 99.01 | 33.76 | 50.35 | 94.55 | 82.84 | 88.31 |
| E | Names+Desig+modDic | SL | **81.12** | **60.53** | **69.33** | **99.69** | **35.47** | **52.32** | **99.01** | **33.76** | **50.35** | **95.25** | **83.45** | **88.96** |
| | | IDF | 79.17 | 59.08 | 67.66 | 99.08 | 35.25 | 52.00 | 98.42 | 33.56 | 50.05 | 93.60 | 82.01 | 87.42 |
| | | SLIDF | 81.02 | 60.45 | 69.24 | **99.69** | **35.47** | **52.32** | **99.01** | **33.76** | **50.35** | **95.25** | **83.45** | **88.96** |
| **LaBSE** | | | | | | | | | | | | | | |
| A | Baseline | SL | 87.36 | 65.19 | 74.66 | 99.50 | 35.57 | 52.41 | 99.41 | 33.89 | 50.55 | **99.11** | **86.84** | **92.57** |
| | | IDF | 87.46 | 65.26 | 74.75 | **99.97** | **35.60** | **52.50** | 99.41 | 33.89 | 50.55 | 98.99 | 86.73 | 92.45 |
| | | SLIDF | 87.36 | 65.19 | 74.66 | **99.97** | **35.60** | **52.50** | 99.41 | 33.89 | 50.55 | **99.11** | **86.84** | **92.57** |
| B | Names | SL | 87.06 | 64.96 | 74.40 | 99.50 | 35.57 | 52.41 | **99.51** | **33.93** | **50.61** | **99.11** | **86.84** | **92.57** |
| | | IDF | 87.36 | 65.19 | 74.66 | 99.18 | 35.28 | 52.05 | 99.41 | 33.89 | 50.55 | 98.48 | 86.29 | 91.98 |
| | | SLIDF | 87.26 | 65.11 | 74.58 | 99.50 | 35.57 | 52.41 | **99.51** | **33.93** | **50.61** | **99.11** | **86.84** | **92.57** |
| C | Names+Desig | SL | 87.06 | 64.96 | 74.40 | 99.50 | 35.57 | 52.41 | **99.51** | **33.93** | **50.61** | **99.11** | **86.84** | **92.57** |
| | | IDF | 87.46 | 65.26 | 74.75 | **99.97** | **35.60** | **52.50** | **99.51** | **33.93** | **50.61** | 98.99 | 86.73 | 92.45 |
| | | SLIDF | 87.26 | 65.11 | 74.58 | 99.50 | 35.57 | 52.41 | **99.51** | **33.93** | **50.61** | **99.11** | **86.84** | **92.57** |
| D | Names+Desig+Dic | SL | 87.46 | 65.26 | 74.75 | **99.97** | **35.60** | **52.50** | **99.51** | **33.93** | 50.60 | **99.11** | **86.84** | **92.57** |
| | | IDF | 87.56 | 65.34 | 74.83 | 99.90 | 35.54 | 52.43 | **99.51** | **33.93** | 50.60 | 98.48 | 86.28 | 91.98 |
| | | SLIDF | 87.51 | 65.30 | 74.79 | **99.97** | **35.60** | **52.50** | **99.51** | **33.93** | 50.60 | **99.11** | **86.84** | **92.57** |
| E | Names+Desig+modDic | SL | **87.71** | **65.45** | **74.96** | **99.97** | **35.60** | **52.50** | **99.51** | **33.93** | 50.60 | **99.11** | **86.84** | **92.57** |
| | | IDF | 86.86 | 64.82 | 74.24 | 99.69 | 35.47 | 52.32 | **99.51** | **33.93** | 50.60 | 98.61 | 86.40 | 92.10 |
| | | SLIDF | 87.66 | 65.41 | 74.92 | **99.97** | **35.60** | 52.50 | 99.31 | 33.86 | 50.50 | **99.11** | **86.84** | **92.57** |

LASER2 embeddings. This improvement was consistently observed for most of the web sources as well. The army news source performed well even from the baseline for the English-Sinhala language pair, owing to the high correlation in the document content. Thus, gains on this data source are not visible. Moreover, the highest gain was observed when using all the bilingual lists.

Unlike English-Sinhala, LASER2 baseline for English-Tamil and Sinhala-Tamil showed very low results. The reason for this may be due to the under-representation of the Tamiltraining data when pre-training the LASER2 model, irrespective of the language's agglutinative nature. However, using the scoring function improvement, we were able to gain a significant improvement on F1 of 10%-20% when using LASER2. The gains for XLM-R and LaBSE are around 1%-3%. We believe that these two multiPLMs, although trained with monolingual data had been capable of capturing the cross-lingual features better than LASER2. This is because they have been trained with massive amounts of data and with a much powerful Transformer[8] architecture with more parameters. Thus, the amount of additional cross-lingual information that comes from the use of bilingual lexicons is less compared to LASER2.

Even though the person names bilingual list of Sinhala-Tamil is about ten times larger than that for the other language pairs, we could not see a considerable improvement in Sinhala-Tamilcompared to the other two. This may be due to the inflected nature of the two languages. The names could be in the inflected form in the parallel content, while the lexicons contain the names in the base form.

XLM-R and LaBSE baseline scores were superior to the LASER2 scores for all three language pairs, which means the multiPLMs undergone pre-training using a massive collection of monolingual data produce improved embeddings even for low-resource languages. The gains for English-Tamil and Sinhala-Tamil language pairs were significant, compared to English-Sinhala. This further asserts that non-English centric language pairs such as Sinhala-Tamil greatly benefit from multiPLMs.

### 4.4.2 Sentence Alignment

For the sentence alignment experiments, we used four baselines.:

- **LASER2**: As mentioned in Section 4.3.4, they used LASER2 multilingual embeddings with margin-based cosine similarity and considered the alignments based on Forward, Backward and Intersection criteria. We recreate this baseline using our dataset.

- **Hunalign**: To compare our work with a statistical method, we recreated this baseline. Hunalign (Varga et al., 2007) had been considered as a baseline in other sentence alignment works as well (Bañón et al., 2020).

[8]LASER2 is built on the RNN architecture

- **LaBSE**: We create this baseline using LaBSE embeddings for our dataset.
- **LaBSE_cos**: Feng et al. (2022) had used raw cosine similarity in their work for the sentence alignment task. We create this baseline to compare the effect of the margin-based cosine similarity over raw cosine scoring measurements.

We conduct the subsequent experiments by incorporating bilingual lists with the improved scoring function proposed by Rajitha et al. (2020) using embeddings obtained from the three multiPLMs.

As our ground truth alignment contained only a small fraction (approx. 300) of parallel sentences, there can be many more valid cross-lingual sentence pairs in these datasets. Therefore we evaluated aligned sentence pairs using Recall (i.e. what percentage of the sentence pairs in the golden alignment set are found by the algorithm), which was one of the commonly used measurements in other research as well (Thompson and Koehn, 2019; Artetxe and Schwenk, 2019a). The results are shown in Table 4.9. Note that, since the use of all the bilingual dictionaries gave the best result for the document alignment task, here we considered only that scenario.

Table 4.9: Sentence Alignment Results in terms of Recall (R). Here, *B* refers to the score obtained using Artetxe and Schwenk (2019a)'s method and *B+D* refers to the scores obtained using Rajitha et al. (2020) bilingual lexicon improvement.

| | Exps | Army | | | | | | Hiru | | | | | | ITN | | | | | | Newsfirst | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | FW | | BW | | I | | FW | | BW | | I | | FW | | BW | | IN | | FW | | BW | | IN | |
| | | Sents | R | Sents | R | Sents | R | Sents | R | Sents | R | Sents | R | Sents | R | Sents | R | Sents | R | Sents | R | Sents | R | Sents | R |
| **Sinhala-English** | | | | | | | | | | | | | | | | | | | | | | | | | |
| Hugalign | | 4352 | | | | | 29.00 | 1650 | | | | | 11.63 | 688 | | | | | 9.00 | 576 | | | | | 11.00 |
| LaBSE_cos | B | 11202 | 98.67 | 12385 | 97.34 | 10145 | 98.33 | 8148 | **97.34** | 6621 | **97.34** | 4757 | **97.34** | 2452 | 99.33 | 2535 | **99.33** | 2535 | 99.00 | 2045 | 98.33 | 1844 | 98.67 | 1268 | **98.33** |
| | B+D | 11202 | 98.00 | 12385 | 98.00 | 10145 | 97.00 | 8148 | **97.34** | 6621 | **97.34** | 4713 | **97.34** | 2452 | 98.33 | 2535 | 99.00 | 1722 | 99.00 | 2045 | 98.33 | 1844 | **99.00** | 1268 | **98.33** |
| LASER2 | B | 11202 | 94.33 | 12385 | 97.00 | 9817 | 93.33 | 8148 | 95.35 | 6621 | 95.35 | 4672 | 94.02 | 2452 | 93.33 | 2535 | **93.33** | 1673 | 89.33 | 2045 | 95.67 | 1844 | 94.33 | 1277 | 92.33 |
| | B+D | 11202 | 96.33 | 12385 | 97.33 | 9901 | 94.33 | 8148 | 95.68 | 6621 | 95.68 | 4806 | 94.35 | 2452 | 95.33 | 2535 | 97.00 | 1724 | 94.00 | 2045 | 97.67 | 1844 | 96.33 | 1263 | 96.00 |
| XLM-R | B | 11202 | 92.33 | 12385 | 93.33 | 9719 | 89.67 | 8148 | 96.35 | 6621 | 96.68 | 4919 | 95.68 | 2452 | 94.00 | 2535 | 96.00 | 1756 | 92.33 | 2045 | 96.67 | 1844 | 95.33 | 1332 | 94.33 |
| | B+D | 11202 | 96.00 | 12385 | 94.67 | 9973 | 93.00 | 8148 | **97.34** | 6621 | 96.68 | 4970 | 96.68 | 2452 | 96.67 | 2535 | 96.67 | 1790 | 96.00 | 2045 | 97.33 | 1844 | 96.67 | 1346 | 95.67 |
| LaBSE | B | 11202 | **99.00** | 12385 | **99.33** | 10340 | **99.00** | 8148 | **97.34** | 6621 | **97.34** | 5114 | **97.34** | 2452 | **99.67** | 2535 | **99.33** | 1854 | **99.33** | 2045 | 98.33 | 1844 | 98.67 | 1376 | **98.33** |
| | B+D | 11202 | **99.00** | 12385 | **99.33** | 10330 | **99.00** | 8148 | **97.34** | 6621 | **97.34** | 5109 | **97.34** | 2452 | **99.67** | 2535 | **99.33** | 1854 | **99.33** | 2045 | **98.67** | 1844 | **99.00** | 1372 | **98.33** |
| **Tamil-English** | | | | | | | | | | | | | | | | | | | | | | | | | |
| LaBSE_cos | BL | 9949 | 94.67 | 10919 | 93.33 | 7855 | 89.33 | 5447 | 88.67 | 4929 | 85.33 | 2979 | 80.33 | 845 | 90.60 | 809 | 90.60 | 514 | 89.60 | 2001 | 96.00 | 1949 | 95.67 | 1414 | 95.67 |
| | B+D | 9949 | 93.33 | 10919 | 94.67 | 8336 | 92.33 | 5447 | 88.00 | 4929 | **86.33** | 3324 | 81.67 | 845 | 91.61 | 809 | **91.95** | 578 | 88.59 | 2001 | 95.67 | 1949 | **96.33** | 1409 | 94.67 |
| LASER2 | B | 9949 | 77.33 | 10919 | 73.67 | 6146 | 67.67 | 5447 | 68.00 | 4929 | 52.00 | 2394 | 44.33 | 845 | 67.11 | 809 | 62.75 | 403 | 54.03 | 2001 | 74.33 | 1949 | 65.33 | 982 | 60.33 |
| | B+D | 9949 | 84.67 | 10919 | 80.67 | 6791 | 76.00 | 5447 | 78.67 | 4929 | 61.67 | 2635 | 56.33 | 845 | 80.54 | 809 | 73.83 | 452 | 69.13 | 2001 | 85.33 | 1949 | 76.00 | 1106 | 73.67 |
| XLM-R | B | 9949 | 86.67 | 10919 | 88.33 | 7531 | 82.00 | 5447 | 83.00 | 4929 | 78.33 | 3235 | 72.67 | 845 | 83.22 | 809 | 83.56 | 537 | 78.86 | 2001 | 92.33 | 1949 | 91.33 | 1340 | 89.33 |
| | B+D | 9949 | 88.33 | 10919 | 91.33 | 7777 | 84.00 | 5447 | 83.67 | 4929 | 79.67 | 3284 | 74.67 | 845 | 85.91 | 809 | 84.56 | 550 | 82.22 | 2001 | 92.67 | 1949 | 93.00 | 1363 | 91.00 |
| LaBSE | B | 9949 | **96.33** | 10919 | 96.33 | 8342 | 94.67 | 5447 | **89.67** | 4929 | **86.33** | 3359 | **83.33** | 845 | **92.62** | 809 | **91.95** | 584 | **91.28** | 2001 | 96.33 | 1949 | **96.33** | 1414 | 96.00 |
| | B+D | 9949 | **96.33** | 10919 | **97.00** | 8336 | **95.33** | 5447 | 88.33 | 4929 | **86.33** | 3324 | 82.33 | 845 | 92.28 | 809 | 91.61 | 578 | 90.60 | 2001 | **96.67** | 1949 | **96.33** | 1409 | **96.33** |
| **Sinhala-Tamil** | | | | | | | | | | | | | | | | | | | | | | | | | |
| LaBSE_cos | BL | 9239 | 93.38 | 9128 | 93.38 | 6112 | 90.73 | 10481 | 93.38 | 10143 | 93.38 | 6048 | 90.73 | 568 | 97.00 | 578 | 97.33 | 445 | 95.67 | 753 | 96.00 | 793 | 97.33 | 540 | 93.33 |
| | B+D | 9239 | 93.71 | 9128 | 91.72 | 6682 | 89.73 | 10481 | 96.33 | 10143 | 97.00 | 6048 | 94.67 | 568 | 99.00 | 578 | 95.67 | 415 | 95.00 | 753 | 97.00 | 793 | 96.33 | 548 | **96.67** |
| LASER2 | B | 9239 | 71.52 | 9128 | 79.47 | 5745 | 66.56 | 10481 | 75.00 | 10143 | 80.33 | 5129 | 69.00 | 568 | 73.00 | 578 | 81.33 | 338 | 65.33 | 753 | 73.00 | 793 | 83.00 | 440 | 67.33 |
| | B+D | 9239 | 74.50 | 9128 | 81.46 | 5920 | 69.54 | 10481 | 80.33 | 10143 | 88.00 | 5314 | 76.00 | 568 | 81.33 | 578 | 86.00 | 351 | 71.67 | 753 | 78.00 | 793 | 88.67 | 466 | 70.33 |
| XLM-R | B | 9239 | 83.44 | 9128 | 81.46 | 6502 | 78.15 | 10481 | 90.67 | 10143 | 91.00 | 6531 | 87.33 | 568 | 91.33 | 578 | 90.00 | 412 | 87.00 | 753 | 93.67 | 793 | 95.33 | 544 | 92.33 |
| | B+D | 9239 | 86.09 | 9128 | 82.45 | 6642 | 79.47 | 10481 | 92.00 | 10143 | 94.33 | 6515 | 91.00 | 568 | 93.33 | 578 | 93.33 | 415 | 90.00 | 753 | 95.00 | 793 | **98.67** | 548 | 93.00 |
| LaBSE | B | 9239 | **95.03** | 9128 | **94.70** | 7162 | **92.38** | 10481 | 97.33 | 10143 | **98.00** | 6679 | 96.33 | 568 | 99.33 | 578 | **98.67** | 445 | **98.33** | 753 | **98.67** | 793 | 97.67 | 567 | 95.33 |
| | B+D | 9239 | **95.03** | 9128 | **94.70** | 7155 | **92.38** | 10481 | **97.67** | 10143 | **98.00** | 6722 | **97.00** | 568 | **99.67** | 578 | **98.67** | 443 | **98.33** | 753 | 97.00 | 793 | **98.67** | 569 | **96.67** |

First and foremost, we note that multilingual embedding-based methods significantly outperform the statistical method Hunalign. Even the baseline Artetxe and Schwenk (2019a), outperforms Hunalign by a significant margin of 74% recall points.

Compared to the LaBSE baseline that uses raw cosine similarity, using the margin-based cosine similarity reports a recall score that is approximately 3% higher for the `Sinhala-Tamil` and `Tamil-English` pairs. Therefore, we can conclude that margin-based cosine is similarly favourable for the sentence alignment task.

Our sentence alignment system outperforms the method in all three language pairs for all the websites with the exception of very few as seen in Table 4.9. Tamil-English language pair shows the highest improvement by outperforming the baseline system by on average 15% recall. For Sinhala-Tamil and Sinhala-English pairs, on average 8% and 4% recall gains (respectively) were obtained for LASER2 embeddings. For XLM-R and LaBSE embeddings, the recall gain was around 3%.

The baseline scores for Tamil-English and Sinhala-Tamil language pairs are considerably low compared to Sinhala-English for LASER2 embeddings. The low amount of training data used for Sinhala and Tamil when training the LASER2 toolkit could be the reason for that. Language diversity and the different forms of inflectional nature of Sinhala and Tamil may also have contributed to this problem. When considering the XLM-R and LaBSE, it was noted that the baseline systems outperform the supervised LASER2 embedding scores. They perform quite well for English-Tamil and Sinhala-Tamil languages as well. We believe the XLM-R and LaBSE embeddings, despite being trained in an unsupervised manner just with monolingual data, have captured the cross-lingual information better than LASER2, as in the document alignment task.

### 4.4.3 Extrinsic Evaluation with NMT

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

NMT systems fine-tuned on the mBART50[9] pre-trained model (Liu et al., 2020) had been successful in terms of Sinhala and Tamil (Thillainathan et al., 2021; Lee et al., 2022). Therefore, in order to build an NMT model, we decided to fine-tune the mBART50 model with the parallel sentences obtained from the sentence alignment task. Experiments were done using the fairseq toolkit (Ott et al., 2019), and the performance was evaluated using the evaluation datasets mentioned in Section 4.3.1. BLEU scores were obtained using sacreBLEU (Post, 2018a) library.

The NMT results shown in Table 4.10 are rather low, which we believe is due to the following reasons: (1): the SiTa evaluation dataset has been obtained from the official document domain while the Flores evaluation datasets have been obtained from Wikipedia. In contrast, we mined the parallel corpus from the news domain. Therefore the domain difference is identified as the primary reason for the NMT systems to produce low results. (2) the parallel corpus size produced by the sentence alignment task is in the range of 9,000-23000, which marks an extremely low-resource setting (Ranathunga

[9] https://huggingface.co/facebook/mbart-large-50

Table 4.10: BLEU Scores for NMT systems trained with parallel data obtained from Sentence Alignment step with Forward (F), Backward (B) and Intersection (I) criterion

| multiPLM | Exp | F | | B | | I | | F | | B | | I | | F | B | I |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Si→En | | | | | | Ta→En | | | | | | Si→Ta | | |
| | | SiTa | Flores | SiTa | Flores | SiTa | Flores | SiTa | Flores | SiTa | Flores | SiTa | Flores | SiTa | SiTa | SiTa |
| LASER2 | BL | 9.7 | 3.9 | 11.6 | 5.6 | 12.0 | 6.3 | 3.8 | 2.1 | 6.4 | 4.1 | 6.6 | 4.8 | 3.5 | 4.4 | 4.5 |
| | BL+Dict | **9.9** | **4.4** | 12.2 | **6.6** | **12.4** | 6.4 | **5.5** | 4.3 | 7.7 | 5.5 | 7.3 | 5.1 | 3.8 | **4.9** | 4.6 |
| XLM-R | BL | 8.8 | 4.0 | 11.4 | 5.6 | 11.9 | 6.5 | 4.0 | 4.1 | 6.1 | 5.1 | 7.7 | 5.9 | 3.7 | 4.1 | 4.7 |
| | BL+Dict | 9.0 | 3.6 | 11.8 | 6.0 | 12.1 | 6.4 | 4.6 | **5.5** | 7.0 | 5.4 | 7.7 | 5.8 | 3.9 | 4.7 | 4.6 |
| LaBSE | BL | 9.5 | 4.3 | 11.9 | 6.3 | 11.9 | **6.6** | 3.8 | 4.4 | 8.1 | 5.8 | 8.2 | 6.2 | **4.0** | 4.7 | 4.7 |
| | BL+Dict | 9.3 | 4.1 | **12.4** | 6.5 | 12.1 | **6.6** | 3.9 | 5.4 | **8.2** | **6.3** | **8.4** | **6.5** | **4.0** | 5.2 | **4.9** |
| | | En→Si | | | | | | En→Ta | | | | | | Ta→Si | | |
| LASER2 | BL | **8.3** | **1.8** | 6.5 | 0.6 | 8.5 | 1.4 | 4.5 | 1.3 | 3.8 | 0.5 | 4.4 | 0.7 | 4.8 | 3.1 | 6.4 |
| | BL+Dict | **8.3** | 1.6 | 6.9 | 0.5 | 8.6 | 1.5 | 4.5 | **1.5** | 4.1 | 0.7 | 4.4 | 1.1 | 6.6 | 3.3 | **6.5** |
| XLM-R | BL | 8.0 | 1.7 | 7.0 | 0.6 | 7.9 | 1.7 | 4.6 | 1.3 | 4.2 | 0.9 | 4.4 | **1.4** | 5.6 | **4.6** | 6.1 |
| | BL+Dict | 8.1 | **1.8** | **7.9** | **0.8** | 8.3 | 1.8 | **4.7** | 1.4 | 4.1 | 0.9 | 4.5 | 1.3 | 5.9 | 4.3 | 5.7 |
| LaBSE | BL | 8.2 | 1.7 | 7.4 | **0.8** | 8.2 | **2.0** | **4.7** | 1.1 | **4.3** | 0.8 | 4.6 | 1.2 | **6.9** | 4.4 | 6.1 |
| | BL+Dict | 8.2 | 1.7 | 7.2 | **0.8** | **8.7** | 1.9 | 4.5 | 1.4 | 4.2 | **1.0** | **5.0** | **1.4** | 5.9 | 4.3 | 6.4 |

et al., 2021). Both these reasons lead for the NMT system to produce a low result. However, we believe that this is not a bottleneck in conducting our study as we are only interested in analysing the impact of the bilingual lexicon integration on the sentence alignment task.

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

We see that the NMT scores obtained by bilingual lists have improved over the baseline scores for most of the cases, as per Table 4.10. This means that bilingual list integration has improved the quality of the parallel sentences. Considering the SiTa evaluation set, the maximum gain provided for LASER2 is +1.8 BLEU, XLM-R is +0.9 BLEU and for LaBSE it is +0.5 BLEU. Similarly, for Flores evaluation set, it is +1.4, +0.6 and +0.5 BLEU for LASER2, XLM-R and LaBSE (respectively). Here we can see identical patterns with respect to both evaluation sets - the gain is the highest for

LASER2, while for XLM-R and LaBSE it is in the same range. Although the Wikipedia data has been used during training these multilingual multiPLMs, it is evident that the multilingual embeddings are not biased to the evaluation set on Wikipedia.

For Ta→En and Ta→Si directions, it shows a maximum improvement of +1.7 BLEU and +1.3 BLEU scores (respectively) for the LASER2 embeddings for the SiTa evaluation set. As Tamil is an under-represented language in the LASER2 training data, the lexicon integration has managed to improve the NMT scores.

In sentence alignment results, the scores were always in increasing order for LASER2, XLM-R and LaBSE respectively. However, for the downstream NMT task, we observed that the scores were mostly high for LASER2 and LaBSE compared to XLM-R. Although we expected the sentence alignment scores and NMT scores to follow the same pattern, it was not the case. The LASER2 had been trained purely on parallel data while LaBSE had been pre-trained using monolingual and parallel data, followed by a fine-tuning phase with parallel data. Therefore, we observe that multilingual systems pre-trained with parallel data perform better in the NMT downstream task.

## 4.5 Discussion

We conducted further analysis to identify the impact of lexicon integration on the sentence alignment task. Table 4.11 shows three scenarios where lexicon integration did not work. An example is given from the Sinhala-English pair. However, these findings are valid for other language pairs as well.

As explained in scenario A, the sentence pair that should be aligned does not contain any overlapping terms with the bilingual lexicons. Hence, such sentences cannot benefit from the dictionary-based improved scoring mechanism. Further, the En sentence and another Si sentence from the same context have overlaps in terms of parallel lexicons. As a result, the sentence alignment algorithm selects an incorrect Sinhala sentence as the alignment for the English source sentence. This shows the bias introduced due to the dictionary-based improvement.

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

According to scenario C, the sentences contain lexicon terms, but in an inflected form. Thus, our algorithms cannot identify those lexical terms appearing in sentences. In the example, the lexicon overlaps are missed for two word-pairs owing to inflections (in both En and Si). If the inflections were accounted in the algorithm, due to the high overlap, the correct alignment sentence pair would be identified. We believe that if a matching can be done at the lemma, a further improvement can be obtained. However,

Table 4.11: Error Analysis in the sentence alignment task. Here, the alignment[corr] refers to the alignment in the gold-standard evaluation set and alignment[incorr] refers to the alignment produced in the experiments.

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

for Sinhala and Tamil, there is no lemmatizer that guarantees the coverage of the full vocabulary. Therefore, at present, working at the lemma level is not feasible.

## 4.6 Chapter Summary

In this chapter, we addressed our second research objective, **RO2: Conduct an empirical Study to determine the impact of different characteristics of the Pre-trained Multilingual Language Models on the Document Alignment and Sentence Alignment tasks for LRLs**.

Our findings indicate that leveraging embeddings from a multiPLM that has undergone unsupervised representation learning followed by fine-tuning on parallel data substantially enhances performance in both document and sentence alignment tasks. During the fine-tuning stage, using parallel data provides direct cross-lingual supervision, allowing the model to better align semantically equivalent sentences between the languages. This is particularly evident in the performance gains achieved by LaBSE compared to LASER2 and XLM-R. Furthermore, incorporating an improved scoring function yields notable improvements for both alignment tasks when used with LASER2 embeddings; however, the effect is less pronounced with LaBSE. We hypothesize that this is due to LaBSE being explicitly fine-tuned for sentence retrieval, enabling it to capture cross-lingual semantics more effectively than LASER2. Apart from the findings, in our empirical study, we release human-annotated, gold-standard benchmark evaluation sets for document and sentence alignment, covering three low-resource language pairs: English-Sinhala, English-Tamil and Sinhala-Tamil.

# CHAPTER 5

# LINGUISTIC ENTITY MASKING (LEM)

## 5.1 Introduction

In the previous chapter, we empirically proved that the embeddings obtained from Pre-trained Language Models (multiPLMs) trained with parallel data is favourable for the document alignment and sentence alignment subtasks. The enhancement can be attributed to parallel data providing explicit cross-lingual signal for the multiPLMs to produce embeddings in close proximity for semantically equivalent sentences in the multilingual space. In this chapter, we explore how to improve the unsupervised representations learnt by an existing multiPLM to better perform for cross-lingual sentence retrieval tasks. Thereby we aim to use the enhanced representations to improve the sentence alignment task.

Encoder-based multiPLMs, such as mBERT (Devlin et al., 2019a) and XLM-R (Conneau et al., 2020a), do not have an explicit cross-lingual pre-trained objective (Hu et al., 2021b) to produce optimal results for sentence retrieval tasks such as sentence alignment. To address this limitation, the Translation Language Modelling (TLM) objective in XLM (Conneau and Lample, 2019) was introduced to enhance the cross-lingual capabilities of existing multiPLMs by leveraging parallel data in a continual pre-training step. However, both MLM and TLM training objectives primarily focus on token-level reconstruction, where the tokens are selected randomly. In this study, we propose a more linguistically informed masking strategy to improve the cross-lingual alignment between the embeddings produced by a multiPLM in a continual pre-training step.

We conduct comprehensive experiments aimed at addressing several key aspects:

- To identify the most effective type of monolingual data for the initial continual pre-training step, specifically comparing **dependent monolingual data** (where source and target sides of parallel data are treated separately as monolingual data) with **independent monolingual data** (monolingual data that does not contain any explicit translation relationship between the two languages)

- To empirically evaluate the performance of existing masking strategies, including sub-word masking (Devlin et al., 2019a), whole-word masking (Devlin et al., 2019a), and span masking (Joshi et al., 2020) for the sentence alignment task

- To determine the most influential linguistic entity or combination of entities for masking.

- To identify the optimal number of tokens to be masked within a linguistic entity.

- To assess the impact of incorporating noisy parallel sentences during the continual pre-training phase.

We address our third research objective in this chapter, **RO3. Improve the cross-lingual representations of existing multiPLMs to obtain High-Quality parallel sentences from the parallel sentence alignment task.**

## 5.2 Motivation

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 (Nastase and Merlo, 2024, 2023; Aoyama and Schneider, 2022) sucessfully. 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 ජැක් පාර දෙසට ගමන් කරනවා in Figure 5.1. In the English sentence, the words "Jack" (NE) and "walk" (Verb) get the highest attention from other words. Similarly, the words "ජැක්" (NE) and "පාර" (Noun) get the highest attention in the Sinhala sentence.

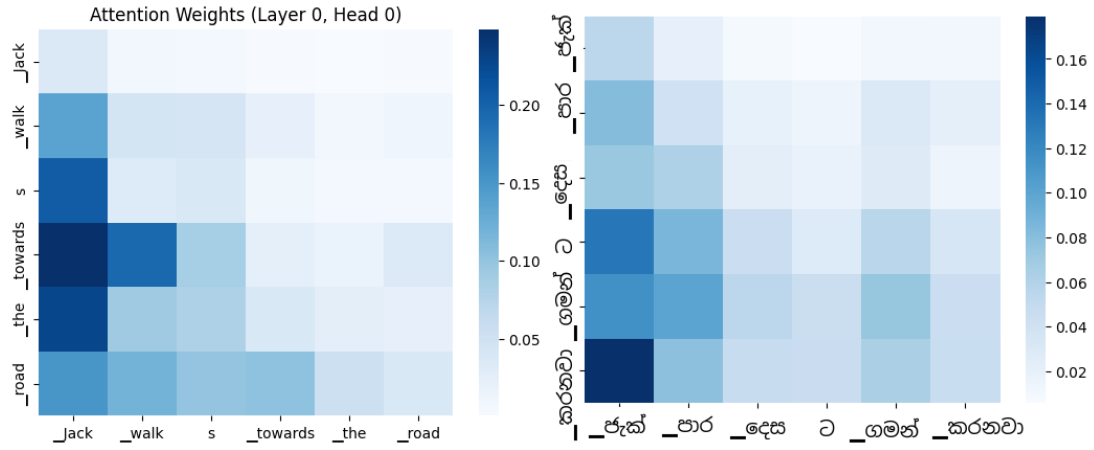


Figure 5.1: Self-attention weights among the words for an English and its corresponding Sinhala sentence. The darker the colour is, the stronger the relationship (ie. self-attention weight) between the two words.

Based on this hypothesis, we propose a linguistically driven masking strategy called *Linguistic Entity Masking (LEM)*. LEM involves masking only a single token from the span of a linguistic entity. We define linguistic entities as **NEs, nouns, and verbs**. This

single-token masking approach differs from existing span masking methods (Sun et al., 2019; Joshi et al., 2020; Levine et al., 2020), as they focus on masking consecutive tokens from selected n-gram spans. We implement LEM leveraging both monolingual sentences ($LEM_{mono}$, comparable to MLM) and parallel sentences ($LEM_{para}$, comparable to TLM), in a continual pre-training an existing multiPLM.

## 5.3 Related Work

### 5.3.1 MLM and TLM Objectives

Encoder-based multiPLMs such as mBERT and XLM-R were trained on monolingual data using the MLM objective. These models have significantly enhanced the performance of various downstream tasks (Rajpurkar et al., 2016; Lai et al., 2017; Wang et al., 2018a).

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 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 sentence alignment, due to the lack of an explicit objective for improving cross-lingual representations (Hu et al., 2021b).

To address this limitation, Conneau and Lample (2019) introduced Translation Language Modeling (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, thereby 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.

### 5.3.2 Different Masking Strategies

BERT's MLM strategy involves masking sub-words. Subsequent research has explored various token masking approaches, as summarized in Table 5.1. Joshi et al. (2020)

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

focused on masking consecutive sub-words within text spans, while Levine et al. (2020) employed Point-wise Mutual Information (PMI) masking to identify and mask correlated text spans. Additionally Golchin et al. (2023) proposed an in-domain keyword masking technique aimed at domain adaptation of PLMs. Notably, most of these techniques have been developed using monolingual data and have been limited to evaluating on high-resource languages.

The most similar work to ours is Entity/Phrase masking (Sun et al., 2019), but it differs from our approach in three key aspects. First, Entity/Phrase masking selects named entities (NEs), noun phrases, and verb phrases for masking. In contrast, as shown in Fig 5.1, we focus exclusively on verb and noun words along with NEs for masking. Second, while Entity/Phrase masking masks all consecutive tokens within NEs or noun/verb phrases identified by a chunking tool, LEM strategy adopts a more targeted approach by masking only a single token within the span of the selected linguistic entity. Lastly, Entity/Phrase masking follows a multi-staged continual pre-training process consisting of sub-word masking (similar to BERT), followed by phrase masking and named entity masking. In contrast, our method employs two continual pre-training stages using the same LEM strategy with both monolingual and parallel data. Furthermore, their approach has only been evaluated on high-resource languages (English and Chinese) and has not been extended to incorporate parallel data for cross-lingual enhancement.

Table 5.1: Existing masking strategies. The *Masked Token Type* indicates the type of words considered for masking. We include our masking strategy (LEM) for comparison purposes.

| **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 Sun et al. (2019) | Multi-stage Pre-training | all sub-words in the Named Entity/Noun Phrase |
| Span Masking (spanBERT) Joshi et al. (2020) | Pre-training | all sub-words in the word n-gram span |
| Point-wise Mutual Information (PMI) Masking Levine et al. (2020) | Pre-training | all sub-words in the correlated word-spans |
| Linguistic Entity Masking (ours) | Continual Pre-training | random subword from a linguistic entity |

Wettig et al. (2023) performed an empirical analysis to determine which tokens should be masked and the ideal masking ratios. However, their research was limited to the English language and focused solely on downstream tasks like classification and question-answering only. So far, to the best of our knowledge, no empirical study has specifically investigated these alternative masking strategies in the context of sentence retrieval tasks, particularly for LRLs.

## 5.4 Methodology

In this section, we discuss the LEM strategy in detail. A comparison between our masking strategy and existing masking strategies is presented in Figure 5.2. Instead of

pre-training a multiPLM from scratch, a computationally expensive process, we utilize LEM in a continual pre-training step. This approach is widely adopted to enhance multiPLMs in terms of representation improvements (Conneau and Lample, 2019; Feng et al., 2022).


Figure 5.2: A comparison of existing masking strategies is presented using an example from the English-Sinhala language pair. Sub-word masking, Whole Word masking, span masking, and $LEM_{mono}$ exclusively utilize monolingual sentences during masking. In contrast, TLM and $LEM_{para}$ apply masking on concatenated parallel sentences. Notably, in both $LEM_{mono}$ and $LEM_{para}$, only a single token from the linguistic entity is masked.

Figure 5.3 illustrates our two-stage continual pre-training process. Following a training sequence similar to the MLM and TLM approach used in XLM, we perform continual pre-training on top of the multiPLM using monolingual data first, followed by parallel data.

## 5.5 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. 5.1.

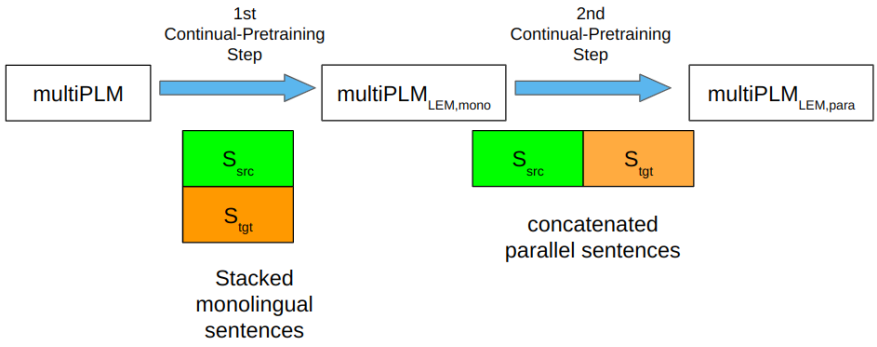


Figure 5.3: The LEM continual pre-training process. An existing *multiPLM*, is first continually pre-trained ( LEM$_{mono}$) with *dependent monolingual* data. In the second continual pre-training step (LEM$_{para}$), the LEM strategy is applied on the *concatenated parallel data*.

.

$$\bar{X} = \bar{x}_1\ \bar{x}_2\ \bar{x}_3\ \bar{x}_4 ..... \bar{x}_j .... \bar{x}_m \tag{5.1}$$

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. 5.2. 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.

$$\bar{X} = \{\{\bar{x}_1\ \bar{x}_2\}, ... \{\bar{x}_4\bar{x}_5\bar{x}_6\}, .....\{\bar{x}_m\}\} \tag{5.2}$$

During training, the cross-entropy loss ($\mathcal{L}_{LEM_{mono}}$) for masked token prediction, as in Eq. 5.3 is minimized. In the equation, N is the total number of tokens for prediction (ie. same as the total number of tokens for masking) and $y_j$ is the expected true label.

$$\mathcal{L}_{LEM_{mono}} = -\frac{1}{N}\sum_{j=1}^{N} y_j \log(P(x_j)) \tag{5.3}$$

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. 5.4. $\bar{X} = \bar{x}_1\ \bar{x}_2\ \bar{x}_3 ..... \bar{x}_k$ 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.

$$\bar{Z} = \bar{x}_1\ \bar{x}_2\ \bar{x}_3 \ldots\ldots\ \bar{x}_k\ \bar{y}_1\ \bar{y}_2\ \bar{y}_3 \ldots\ldots\ \bar{y}_l \tag{5.4}$$

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. 5.5) 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.

$$\mathcal{L}_{LEM_{para}} = -\frac{1}{S}\sum_{s=1}^{S} z_s \log(P(x_s) - \frac{1}{T}\sum_{t=1}^{T} z_t \log(P(j_t)) \tag{5.5}$$

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 5.2 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 Section 5.9.1.

Table 5.2: English (En), Sinhala (Si), and Tamil (Ta) examples of the returned sub-words after the tokenization step are presented. In English, nouns are typically inflected based solely on number. In contrast, Sinhala and Tamil nouns undergo inflection not only based on number but also on case category and gender.

| **Type** | **Singular/Subject** | **Plural/Subject** | **Plural/Object** | **Singular/Feminine** |
|---|---|---|---|---|
| Original word (Si) | ගුරුවරයා | | | |
| Si word (inflected) | ගුරුවරයා | ගුරුවරු | ගුරුවරුන් | ගුරුවරිය |
| Tokenized output | ගුරු #වර #යා | ගුරු #වරු | ගුරු #වරු #න් | ගුරු #වරි #ය |
| Original word (En) | **Teacher** | | | |
| En word (Translation) | teacher | teachers | teachers | the teacher |
| Tokenized output | teacher | teacher #s | teacher #s | the teacher |
| Original word (Ta) | **ஆசிரியர்** | | | |
| Ta word (Translation) | ஆசிரியர் | ஆசிரியர்கள் | ஆசிரியர்கள் | ஆசிரியர் |
| Tokenized output | ஆசிரியர் | ஆசிரியர் #கள் | ஆசிரியர் #கள் | ஆசிரியர் |

## 5.6 Experiments

We detail out the experiments conducted in this study.

### 5.6.1 Impact of the type of monolingual data in $LEM_{mono}$

We conduct experiments using both **independent monolingual data** and **dependent monolingual data** to analyze their impact on the continual pre-training step $LEM_{mono}$. Specifically, we sample 60,000 sentences from the MADLAD-400 (Kudugunta et al., 2024) dataset and use all 60,000 available sentences from the SiTa-Trilingual dataset for each language. Subsequently, we continually pre-train XLM-R separately with these datasets using $LEM_{mono}$ and evaluate the performance on the bitext evaluation dataset.

To determine whether increasing the independent training data leads to performance improvements, we repeat the experiment using a sample size of 100,000 sentences from MADLAD-400. Additionally, as an extreme case, we conduct a third experiment with 500,000 sentences for the `Sinhala-Tamil` language pair. In these experiments, the sizes 60,000, 100,000, and 500,000 represent the number of training sentences per language, with the total training set size being twice the specified amount. This evaluation is performed for all three language pairs.

### 5.6.2 Evaluation of Different Masking Strategies

We conduct an empirical evaluation of various masking strategies and assess their performance on the sentence alignment 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.

### 5.6.3 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 Section 5.7.3.

We identify NEs in `English`, `Sinhala` and `Tamil` sentences using an in-house fine-tuned multilingual NER model (Ranathunga et al., 2024b). To detect nouns and verbs in the sequences, we utilize the Flair POS tagger (Akbik et al., 2018) for `English`, the `Sinhala` TnT POS Tagger (Fernando and Ranathunga, 2018; Fernando et al., 2016) for `Sinhala`, and ThamizhiUDp (Sarveswaran and Dias, 2020) for `Tamil`. The Flair POS tagger, which achieved an F1 score of 98.19%, is considered the best model for

English POS tagging. The Sinhala POS tagger, trained using SVM, reported an overall accuracy of 84.68% and an accuracy of 59.86% for tagging unknown words. The Tamil POS tagger, a neural-based model, obtained an F1 score of 93.27%. These models represent the optimal solutions for POS tagging in Sinhala and Tamil.

The initial ablation experiment focuses on masking a single linguistic entity type, such as only NEs, only verbs, or only nouns. Following this, combinations of these linguistic entity types are explored to evaluate their effectiveness.

In our experiments, the notation *100%NE+15%MLM* indicates that priority is given to sampling from NEs. If the selected NEs do not yield enough tokens for masking, the remaining tokens are sampled from the rest of the input. Similarly, when combining multiple linguistic entities, for example, *100%NE+100%VB+15%MLM*, it means that priority is given to sampling tokens for masking from both NEs and verbs.

#### 5.6.4 Evaluation Tasks

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

##### 5.6.4.1 Sentence Alignment

Advancements in sentence alignment techniques can be broadly divided into two categories: (1) improving the semantic similarity distance calculation function between sentence embeddings (Artetxe and Schwenk, 2019a; Fernando et al., 2023), and (2) enhancing cross-lingual sentence representations (Artetxe and Schwenk, 2019b; Yang et al., 2020; Feng et al., 2022). Our LEM strategy specifically targets improvements in the latter.

We extract sentence embeddings from XLM-R$_{LEM}$ as well as from the baseline models and employ the margin-based cosine similarity function (Artetxe and Schwenk, 2019a) to identify parallel sentence pairs. We opt for margin-based cosine similarity over traditional cosine similarity due to its reduced rate of false positives. The identified parallel sentences are then ranked based on their similarity scores. Bitext mining is performed using three criteria: Forward (FW), Backward (BW), and Intersection (IN), following the work of Artetxe and Schwenk (2019a). FW retrieves the target sentence corresponding to each source sentence, BW retrieves the source sentence corresponding to each target sentence, and IN considers the intersection of parallel sentences retrieved by both FW and BW criteria. We evaluate the sentence alignment task on the golden aligned human curated evaluation set (Fernando et al., 2023). In this evaluation set, humans selected 300 sentence pairs from the available candidates. Therefore, we report the results using the recall metric.

#### 5.6.4.2 Parallel Data Curation

Large-scale bitext mining (Schwenk et al., 2021b; Costa-jussà et al., 2022) helps mitigate the parallel data scarcity issue in NMT. However, these mined bitexts are often noisy (Kreutzer et al., 2022a; Ranathunga et al., 2024a), necessitating a parallel data curation step to filter out noisy sentence pairs from the corpus. This curation process involves obtaining sentence representations from a multiPLM and calculating the semantic similarity between each parallel sentence pair using cosine distance (Feng et al., 2022; Costa-jussà et al., 2022). The parallel sentences are then ranked in descending order of similarity, and the top-ranked sentence pairs are selected for training the NMT system.

To further assess the effectiveness of these models with enhanced cross-lingual representations, we carry out a parallel corpus curation task. We then perform an extrinsic evaluation by training NMT systems using the top-ranked sentences from the curated corpus.

First, we rank the parallel sentences based on translation quality using both the baseline models (Section 5.7.3) and the XLM-R$_{LEM}$ models. From the ranked parallel sentence pairs, we select the top 50,000 sentences and train NMT systems for each language pair. The NMT performance is evaluated on the Flores+ devtest set using the sacreBLEU (Post, 2018b), ChrF (Popović, 2015), ChrF++ (Popović, 2017), and spBLEU (Goyal et al., 2022) metrics. We primarily discuss the results using the ChrF++ metric.

## 5.7 Experiment Setup

### 5.7.1 Data Selection

We considered `English-Sinhala`, `English-Tamil` and `Sinhala-Tamil` as the LRLs for our experiments as per the justification in Section 2.6.

**Monolingual and Parallel Data:** As discussed in Section 5.6.1, we performed an ablation study to identify the most suitable type of monolingual data for the first continual pre-training step. We acquired the independent monolingual data from MADLAD-400 (Kudugunta et al., 2024), a large-scale manually audited collection of document-level data containing 3 trillion tokens from Common Crawl[2] across 419 languages. The dataset cover the languages under our study.

For dependent monolingual data, we used the monolingual sides extracted from the SiTa-Trilingual parallel dataset (Fernando et al., 2020). This dataset is a human-curated, gold-standard, three-way parallel corpus between `Sinhala`, `Tamil` and `English` languages, comprising 60,000 training data instances.

[2] https://commoncrawl.org/

We preprocess the MADLAD-400 data and extracted clean sentences for each language as follows: First, we segmented the document-level data into sentences using the NLTK[3] sentence tokenizer. Next, we filtered these sentences using the LID (**L**anguage **ID**entification) model[4]. Afterward, we removed noisy data, including HTML tags, URLs, and sentences containing less than 60% textual content. Additionally, religious texts were excluded through a keyword filter[5]. In contrast, no preprocessing was applied to the SiTa-Trilingual data, as it is already of high quality.

**NLLB/CCAligned Datasets:** For the parallel data curation task, we obtained parallel data from the NLLB (Costa-jussà et al., 2022) and CCAligned (El-Kishky et al., 2020) corpora. Both of these corpora provided parallel data for the three language pairs: `English-Sinhala`, `English-Tamil` and `Sinhala-Tamil`. However, it was known that the parallel data from NLLB and CCAligned for these language pairs were noisy (Ranathunga et al., 2024a).

**ParaCrawl Dataset:** To analyze the performance of the LEM strategy with noisy data, we selected the `English-Sinhala` ParaCrawl (Bañón et al., 2020) dataset. This dataset was a web-mined parallel corpus consisting of 217,412 parallel sentences.

**Trilingual Bitext Mining Evaluation Set:** For the sentence alignment task, we used an existing human-created dataset (Fernando et al., 2023). It consists of trilingual data obtained from four Sri Lankan news sources Army[6], Hiru[7], ITN[8] and Newsfirst[9]. For each news source, there are human-aligned 300 sentence-pairs.

**NMT Evaluation Set:** For the NMT experiments, we used dev and devtest splits from Flores+[10] as the validation and evaluation sets, respectively.

### 5.7.2 MultiPLM Selection

We chose XLM-R as the base multiPLM for our experiments, as other popular multiPLMs like XLM and mBERT do not cover the `Sinhala` language. Although LASER3 and LaBSE models are potential multiPLMs covering the three language-pairs under our study, they have undergone fine-tuning stages which have already been optimized for *cross-lingual* tasks. XLM-R has already demonstrated promising performance in downstream tasks for the low-resource languages considered in this study (Dhananjaya et al., 2022; Rathnayake et al., 2022; Udawatta et al., 2024; Ranathunga et al., 2024b). It is a 278M parameter model, pre-trained on 100 languages. However, the amount of `Sinhala` and `Tamil` data used during XLM-R pre-training is significantly lower than

[3] https://www.nltk.org/index.html
[4] https://github.com/gordicaleksa/Open-NLLB
[5] Keywords include Bible book names along with common words from the Bible.
[6] https://www.army.lk/
[7] https://www.hirunews.lk/
[8] https://www.itnnews.lk/
[9] https://english.newsfirst.lk/
[10] https://github.com/openlanguagedata/flores

that for English (English: 55B, Sinhala: 243M, Tamil: 595M).

### 5.7.3 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**$_{MLM+TLM}$ Conneau and Lample (2019) - We continually pre-train the XLM-R with MLM+TLM objectives and use the representation improved encoder to obtain embeddings.

### 5.7.4 Implementation and Hyper-parameters

#### 5.7.4.1 Linguistic Entity Masking (LEM)

We customized the MLM training implementation provided by the sentence-transformers[11] library (built on Hugging Face transformers[12]) to support XLM-R tokenization and incorporate the LEM strategy. Each continual pre-training experiment is executed for 60 epochs with early stopping, and the checkpoint with the lowest validation loss is selected as the best-performing model.

The experiments were conducted on an NVIDIA Quadro RTX 6000 GPU with 24GB VRAM. The hyperparameters of the XLM-R [13] model, along with other training parameters used in the continual pre-training experiments, are detailed in Table 5.3.

Table 5.3: 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 $\epsilon$ | 1 e-08 |
| Adam $\beta_1$ | 0.9 |
| Adam $\beta_2$ | 0.99 |

[11] https://www.sbert.net/
[12] https://huggingface.co/docs/transformers/index
[13] https://huggingface.co/FacebookAI/xlm-roberta-base

#### 5.7.4.2 NMT Experiments

We extracted the top-ranked sentences for the NMT experiments and trained a Sentencepiece[14] tokenizer with a vocabulary size of 25,000. Next, we utilized the Fairseq toolkit (Ott et al., 2019) to build and trained a vanilla transformer-based Sequence-to-Sequence NMT model. The experiments were conducted on a NVIDIA Quadro RTX6000 GPU with 24GB VRAM. The hyperparameters and training parameters used during the experiments are presented in Table 5.4. Each experiment was run for 100 epochs, with early stopping applied to prevent overfitting.

Table 5.4: 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 $\beta_1$, Adam $\beta_2$ | 0.9, 0.99 |
| warmup-updates | 4000 |
| warmup-init-lr | 1e-7 |
| learning rate | 1e-3 |
| batch-size | 32 |
| patience | 6 |
| fp16 | True |

#### 5.7.4.3 Improving Continual Pre-training Efficiency

Named Entity Recognition (NER) and Part-of-Speech (POS) tagging during training increased training time drastically. We introduced a pre-processing step to mitigate this issue. Specifically, a dictionary was created to store the linguistic entities, such as named entities, verbs, and nouns, for each sentence. We maintained a sub-word-level mapping in the dictionary, allowing for precise token identification while reducing computational overhead during training. The $LEM_{mono}$ step was continually pre-trained with 112K monolingual sentences and $LEM_{para}$ with 56K parallel concatenated sentences. With the improvement made during pre-processing, the continual pre-training times were approximately 15Hrs-16Hrs and 10Hrs-11Hrs for $LEM_{mono}$ and $LEM_{para}$, respectively.

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

## 5.8 Experimental Results

In this section, we report the results obtained for the experiments while emphasizing the empirical findings.

### 5.8.1 Impact of the type of monolingual data in $LEM_{mono}$

Figure 5.4 shows the sentence alignment results for the $LEM_{mono}$ step using independent and dependent monolingual data. The detailed results are available in Table A.1 in Appendix A.

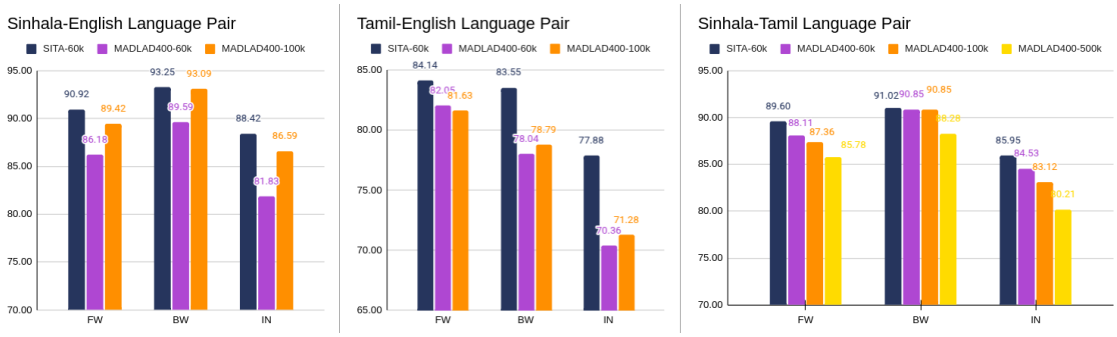


Figure 5.4: Sentence alignment Recall scores for using independent monolingual data (MADLAD-400) versus dependent monolingual data obtained from the parallel corpus (SiTa-Trilingual parallel Corpus). Here the Forward (FW), Backward (BW) and Intersection (IN) approach refers to the criterion followed to identify the translation sentences as per the work of Artetxe and Schwenk (2019a).

The highest performance was consistently observed when using dependent monolingual data across all three language pairs. Interestingly, increasing the size of the independent dataset to 100,000 did not yield better scores compared to those obtained with the dependent monolingual 50,000 training dataset, a trend consistent across all three language pairs. Further increasing the dataset size to 500,000 resulted in even poorer performance.[15]

These findings clearly indicate that utilizing dependent monolingual data during the LEM$_{mono}$ step is beneficial for improving cross-lingual representations.

### 5.8.2 Evaluation of Different Masking Strategies

Table 5.5 presents the experimental results for the sentence alignment task using XLM-R continually pre-trained with various MLM strategies (Section 5.6.2). Based on the averaged recall scores, the baseline XLM-R consistently demonstrated the highest performance across sentence alignment tasks. An exception was observed for the

[15]Due to resource constraints and the diminished performance observed in the Sinhala-Tamil direction, the experiment with 500,000 data points was not conducted for the other two language pairs.

Sinhala-Tamil BW criterion, where the sub-word masking strategy outperformed the baseline. A significant finding from these experiments is that, in general, continual pre-training with existing masking strategies tended to degrade the already acquired cross-lingual representations in the XLM-R model.

As described in Section 5.5, morphologically rich, low-resource languages tend to produce a higher number of sub-word tokens during the tokenization process due to the occurrence of infrequent words within sequences. Consequently, masking strategies such as whole-word masking and span masking result in longer masked spans, thereby reducing the available context for accurate predictions. We hypothesize that this reduction in contextual information lowers prediction accuracy, ultimately weakening the cross-lingual representations already learned by the XLM-R model. As a result, the existing masking strategies yield degraded performance.

Table 5.5: Sentence alignment Recall scores for the different masking strategies.

| 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 |

### 5.8.3 Evaluation of LEM Strategy and Ablation Study

The results of the ablation study, which examine the impact of each linguistic entity and their combinations on the LEM strategy, are presented in Table 5.6 for Sinhala-English, Table 5.7 for Tamil-English and Table 5.8 for Sinhala-Tamil respectively. The final gains are summarized in Table 5.9.

For the Sinhala-Tamil language pair, XLM-R$_{LEM}$ demonstrated the highest gain of +3.1 Recall points compared to the raw embeddings from XLM-R. In comparison with XLM-R$_{MLM+TLM}$, the gain was +1.4 points. For the English-Tamil language pair, the LEM$_{mono}$ step initially produced a lower score compared to the XLM-R baseline. However, after the second continual pre-training step, XLM-R$_{LEM}$ surpassed the XLM-R baseline by +1.2 points. Notably, the highest gain achieved by our method compared to XLM-R$_{MLM+TLM}$ was observed for the English-Tamil language pair, with an improvement of +2.4 points.

Table 5.6: Ablation experiments and sentence alignment scores for English-Sinhala language pair considering linguistic entity masking.

| 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 |
| 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 |
| 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 |
| 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 | 95.33 | 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 5.7: Ablation experiments and sentence alignment scores for English-Tamil language pair considering linguistic entity masking.

| 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 |
| 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 |
| 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 |
| 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 |
| 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 5.8: Ablation experiments and sentence alignment scores for Sinhala-Tamil language pair considering linguistic entity masking.

| 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 |
| 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 |
| 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** |
| 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 |

A similar pattern was observed with the English-Sinhala language pair. Here, XLM-R$_{LEM}$ demonstrated a gain of +0.4 points compared to the baseline XLM-R and an improvement of +1.7 points compared to XLM-R$_{MLM+TLM}$.

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.

Table 5.9: Results for sentence alignment 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 Gain** |
|---|---|---|---|---|
| | **FW** | **BW** | **IN** | |
| **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 |

### 5.8.4 Parallel Data Curation

Table 5.10 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 could 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 produced 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 produced 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 the 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. (2024), 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 were consistent with other metrics, as shown in Table B.1 in the Appendix B as well. The results further proved 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 benefited the parallel data curation task as well.

Table 5.10: 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.

| | **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** |

## 5.9 Ablation Studies

In this section, we describe the ablation experiments conducted to evaluate the LEM strategy further.

### 5.9.1 The Number of Tokens for Masking in LEM Strategy

To assess the impact of the number of masked tokens within linguistic entities, we conducted an ablation study. This study was carried out specifically for the Sinhala-Tamil language pair. As shown in Table 5.8, the best performance was achieved using the *100%NE+15%MLM* combination during the $LEM_{mono}$ step and the *100%NE+15%TLM* combination during the $LEM_{para}$ step.

Based on this optimal setting, we varied the number of tokens to be masked and evaluated the performance on the sentence alignment task. The results, presented in Table 5.11, clearly demonstrate a trend of decreasing performance as the number of masked tokens increases.

When masking only one token per linguistic entity, the average performance across tasks was the highest. This outcome indicates that minimal masking preserves more contextual information, enabling the model to capture dependencies crucial for downstream tasks more effectively. However, as the number of masked tokens increased to

two or more, the average performance significantly declined. The performance drop became particularly pronounced when the token count was raised to three and four. This suggests that excessive masking disrupted the contextual integrity of linguistic entities, resulting in suboptimal representations.

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

Table 5.11: The Recall scores from the ablation study by changing the number of tokens masked in the linguistic entity. The results are for the Sinhala-Tamil language pair and the sentence alignment downstream task.

| No. of Tokens Masked in Linguistic Entity | FW | BW | IN | Average |
|---|---|---|---|---|
| 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 |

### 5.9.2 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[16] dataset.

As per Table 5.6, for `English-Sinhala` best results 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 5.12 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 surpass, and in FW criteria, the scores are the same as those 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.

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

Table 5.12: Sentence alignment Recall results obtained using LEM-enhanced models on both high-quality and noisy web-crawled datasets.

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

## 5.10 Discussion

The LEM strategy is highly dependent on the accuracy of the underlying tools used to identify linguistic entities. Sub-optimal performance of the NER model and POS taggers can adversely affect the final results.

While the NER model demonstrates good performance with `English` sentences, we observe two primary types of errors with `Sinhala` and `Tamil`, as detailed in Table 5.13. The first type of error, classified as False Positives, involves words that are incorrectly tagged as named entities (NEs) despite not being part of an NE. The second error type, False Negatives, occurs when the NER model fails to recognize all words belonging to the NE sequence and incorrectly labels some words.

Table 5.13: Examples of incorrect identification and labeling of NEs. We identify two error categories: false positives and false negatives, where the NER model underperforms.

| | |
|---|---|
| **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 .<br>அரசாங்க **B-ORG** **(Government) is not a NE. Correct POS Tag should be** **NOUN** |
| **False Negatives: NEs, not identified during Correctly** | |
| Si | අපනයන O සංවර්ධන I-MISC සහ O උපදේශන O සේවා O<br>**The entire sentence should be NE therefore the correct tag sequence should be:**<br>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<br>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<br>**The organisation**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<br>திருமதி. O பீ. I-PER கே.எஸ்.எம். I-PER சியாமா I-PER சமரவீர I-PER<br>**Both salutations** திரு. O **(Mr.) and** திருமதி. O **(Mrs.) should be:** B-PER . **Hence**எச். B-PER **should be** I-PER . |

Similar issues are observed with POS tagging, as shown in Table 5.14, where errors

result in both False Positives and False Negatives. However, for the English language, the returned POS tags are generally accurate.

Table 5.14: Examples of incorrect identification and labelling of POS Tags. We identify mainly two error categories: false positives and false negatives, where the Pos Tagger underperforms.

| **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 .<br>**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 | මෑ JJ ඇළ NNC පුනරුත්ථාපනය NNC කිරීම NNC<br>කිරීම NNC **(do) should be a** VERB |
| Si | 10. NUM පොලොන්නරුව NNC මහාධිකරණ NNJ විනිසුරු NNC නිල JJ නිවස ??? ඉදි RPCV කිරීම. NNC<br>නිවස **???** **(house) should be a** NOUN |
| Ta | கொள்வனவு NOUN செய்யப்பட்ட VERB நூல்கள் NOUN அறநெறிப் NOUN பாடசாலை NOUN மாணவர்களுக்கு NOUN விநியோகிக்கப்பட்டன NONE<br>**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 |

## 5.11 Chapter Summary

In this chapter, we address the third research objective, **RO3. Improving the cross-lingual representations of multiPLMs to identify High-Quality parallel sentences for the parallel sentence alignment task**.

We propose a novel masking strategy, Linguistic Entity Masking (LEM), to enhance the cross-lingual representations in existing multiPLMs. The LEM approach involves masking a single token per instance, specifically targeting linguistically informed entities such as NEs, nouns, and verbs. We apply the LEM strategy on XLM-R multiPLM and empirically prove this improvement. When this strategy is further combined with parallel data, it yields even greater improvements, as evident by the performance gains observed across low-resource language pairs English-Sinhala, English-Tamil and Sinhala-Tamilfor sentence alignment and PDC task. We publicly release the cross-lingual improved encoders publicly, to be used by the research community.

# CHAPTER 6

# DEBIASING THE DISPARITY IN NMT SYSTEMS

## 6.1 Introduction

In Chapter 4, we showed that the selected multiPLMs affect the sentence-retrival task performance and in Chapter 5, we showed that the *cross-lingual* representations in existing multiPLMs can be improved to optimize the performance of such tasks further.

In this chapter, we explore the third DA technique (Section 2.5), Parallel Data Curation (PDC). Most large-scale web mined corpora (Costa-jussà et al., 2022; Bañón et al., 2020; Schwenk et al., 2021b) are noisy, especially for LRLs (Bane et al., 2022; Kreutzer et al., 2022b). Therefore, it is important to filter out such noisy parallel sentences prior to training NMT systems (Khayrallah and Koehn, 2018), as the noise adversely affect the translation quality. With PDC, first, the parallel sentences are ranked based on the degree of translation coverage, which is determined based on a semantic similarity score, calculated between the sentence embeddings obtained from a multiPLM. Then the NMT system is trained using the top ranked parallel sentences. However, Ranathunga et al. (2024a); Moon et al. (2023) demonstrated that the choice of multiPLM significantly impacts the ranking process, which results in a disparity among NMT systems trained on the ranked corpus. According to (Moon et al., 2023), this is due to the biases that are inherent to the multiPLMs, which turn out to be noise from a NMT perspective. We hypothesise that this bias can be mitigated with heuristic-based PDC approaches. To investigate this, we conduct a series of ablation experiments to observe the impact of the selected heuristics on the final NMT scores and whether they are capable of minimizing the disparity. Additionally, we conduct a comparative human evaluation to quantify the types of noise that are not filtered due to biases in the multiPLMs during the ranking process.

In this chapter, we address the research objective, **RO4. Exploring Parallel Data Curation (PDC) techniques to extract high-quality parallel sentences from web-mined parallel corpora**.

## 6.2 Motivation

Selecting three multiPLMs, we investigate the disparity among NMT models further. We obtain embeddings from three multiPLMs: LASER3 (Heffernan et al., 2022), XLM-R (Conneau et al., 2020a), and LaBSE (Feng et al., 2022), calculate the semantic similarity between each parallel sentence pair in the CCMatrix Schwenk et al. (2021b) and CCAligned El-Kishky et al. (2020) datasets and rank them in descending order. Then we train NMT models (Section 6.5.4) using the top-ranked 100K sentence pairs

from each corpus. Experiments are carried out for English-Sinhala (En-Si), English-Tamil (En-Ta) and Sinhala-Tamil (Si-Ta). As shown in Figure 6.1, there is a significant disparity among the results, mainly for En-Si and En-Ta language pairs.

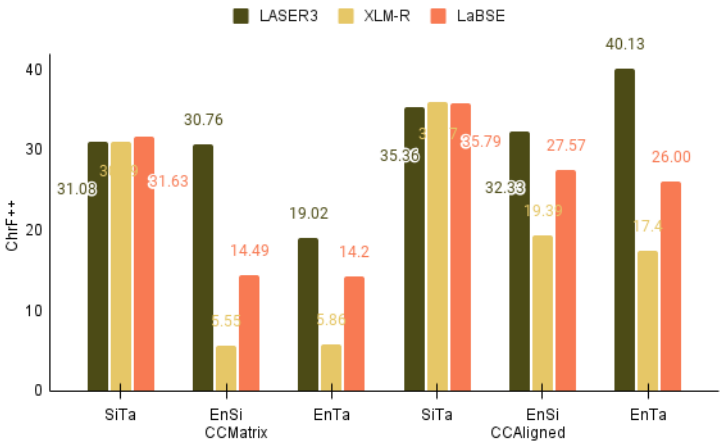


Figure 6.1: Baseline NMT scores in ChrF++ when trained with the top-ranked sentence pairs from CCMatrix and CCAligned, using embeddings obtained from LASER3, XLM-R, and LaBSE.

A manual inspection of the En-Si top-ranked 1000 sentences reveals that different multiPLMs prioritise different sentence characteristics when ranking parallel sentences. For example, sentence pairs ranked top by XLM-R are mostly short and contain overlapping text such as numbers, acronyms, and URLs. LaBSE also favours sentences with numbers and date overlaps, while LASER3 prioratizes relatively better full-length sentence-pairs. The impact of the nature of the top-ranked sentences is evident in Figure 6.1, where results related to LASER3 outperform its two counterparts by a significant margin, in most of the experiments.

In order to carry out a more quantitative evaluation, we randomly selected 100 sentence pairs from the top 1000 sentences from the aforementioned ranked corpora and carried out a human evaluation. Ranathunga et al. (2024a)'s is the most comprehensive error taxonomy available for this task. However, to better capture the noise identified during the manual observation, we extended this error taxonomy (Section 6.4). Human evaluation results in Table 6.2 show that compared to LASER3, the error category percentage (E) is excessively high for XLM-R and LaBSE. The most contributing types of errors are untranslated text (UN), short sentences (CS) and sentences with number/acronym/URL overlaps (CCN). Examples for these are shown in Table 6.1.

We hypothesise that at least some of these noisy sentences can be filtered using rule-based heuristics. Although applying heuristics is a common approach to improving the quality of parallel corpora (Sloto et al., 2023; Steingrímsson et al., 2023), the use of heuristics has not been consistent. They have limited their analysis to a single commonly

Table 6.1: Example parallel sentences from the En-Si, En-Ta and Si-Ta, identified during human evaluation.

| Language Pair | Source Sentence | Target Sentence |
|---|---|---|
| **Incorrect Translation (X) : Both languages correct. But has translation errors** | | |
| En - Si | Ample storage room and slots for credit cards, IDs and Cash | ක්‌රෙඩිට කාඩ, හැඳුනුම් පත් සහ මුදල් සඳහා ඇති තරම් ගබඩා කාමරය සහ මදකව |
| En - Ta | Where will you be in the next five, ten or fifteen years? | கே: அடுத்த ஐந்து, பத்து அல்லது பதினைந்து ஆண்டுகளில் நீங்கள் எங்கே இருப்பீர்கள்? |
| **Untranslated Text (UN): either in source or target side just copied from the translation counterpart** | | |
| En - Si | What do you mean when you say "Your comment is awaiting moderation?" | මොකෝ විචාරක කුමා මගේ කමෙන්ට් එක කාම " Your comment is awaiting moderation. " ? |
| En - Ta | Effective Pixels: 16.0 million (Image processing may reduce the number of effective pixels.) | ஆப்டிகல் சென்சார் ரெசொலூஷன் 20.1 million (Image processing may reduce the number of effective pixels) |
| **Not a language (NL): at least one of source and target are not linguistic content** | | |
| En - Si | 1.5mm2 / 900mm 2.0mm2 / 900mm 2.5mm2 / 900mm 4.0mm2 / 900mm | 1.5mm2 / 900mm 2.0mm2 / 900mm 2.5mm2 / 900mm 4.0mm2 / 900mm |
| En - Ta | HQCCWM750GAH6A | பதிவிறக்க: HQCCWM750GAH6A.pdf |
| **Wrong Language (WL): Source and Target side are linguistic content. However, the source, target, or both sides a not in the expected language.** | | |
| En - Si | Ի պատից պարոն Գոլջի: | ගොල්ගි මහතා විසින් එතුමාගේ නමින් ම නම් කෙරිණි. |
| **Short Sentence (CS): Correct Translation, but the number of tokens on the Source or Target side is less** | | |
| En - Ta | July 21: | ஜூலை 21: |

Table 6.2: Human evaluation results for CCMatrix and CCAligned for En-Si, En-Ta and Si-Ta. Results are reported for LASER3, XLM-R, and LaBSE before and after applying heuristics. We report the average score among the scores obtained from the individual annotators. **(C)** - overall correct percentage considering CC (perfect translation), CN (near perfect) and CB (boilerplate). **(E)** - overall error percentage considering CCN (correct with overlaps), CS (correct but short sentence), X (wrong translation), UN (untranslated), WL (wrong language), NL (not a language).

| | CC | CN | CB | C | CS | CCN | UN | X | WL | E |
|---|---|---|---|---|---|---|---|---|---|---|
| Sinhala-Tamil | | | | | | | | | | |
| LASER3-Before | 8% | 27% | 2% | **37%** | 14% | 14% | 34% | 1% | 0% | **63%** |
| LASER3-After | 16% | 56% | 1% | **73%** | 13% | 4% | 10% | 0% | 0% | **27%** |
| XLM-R-Before | 1% | 10% | 0% | **11%** | 40% | 19% | 29% | 0% | 1% | **89%** |
| XLM-R-After | 0% | 20% | 2% | **22%** | 13% | 29% | 35% | 0% | 1% | **78%** |
| LaBSE-Before | 4% | 6% | 0% | **10%** | 74% | 7% | 9% | 0% | 0% | **90%** |
| LaBSE-After | 29% | 33% | 0% | **62%** | 2% | 32% | 4% | 0% | 0% | **38%** |
| English-Sinhala | | | | | | | | | | |
| LASER3-Before | 17% | 7% | 4% | **28%** | 7% | 10% | 55% | 0% | 0% | **72%** |
| LASER3-After | 39% | 39% | 7% | **85%** | 0% | 7% | 8% | 0% | 0% | **15%** |
| XLM-R-Before | 1% | 0% | 0% | **1%** | 13% | 4% | 80% | 2% | 0% | **99%** |
| XLM-R-After | 3% | 8% | 26% | **37%** | 0% | 2% | 53% | 8% | 0% | **63%** |
| LaBSE-Before | 13% | 2% | 0% | **15%** | 63% | 14% | 8% | 0% | 0% | **85%** |
| LaBSE-After | 87% | 7% | 3% | **97%** | 0% | 1% | 2% | 0% | 0% | **3%** |
| English-Tamil | | | | | | | | | | |
| LASER3-Before | 0% | 3% | 2% | **5%** | 0% | 0% | 95% | 0% | 0% | **95%** |
| LASER3-After | 6% | 61% | 20% | **87%** | 0% | 3% | 10% | 0% | 0% | **13%** |
| XLM-R-Before | 0% | 0% | 2% | **2%** | 3% | 5% | 90% | 0% | 0% | **98%** |
| XLM-R-After | 0% | 39% | 31% | **70%** | 1% | 3% | 21% | 4% | 0% | **30%** |
| LaBSE-Before | 0% | 9% | 2% | **11%** | 34% | 7% | 48% | 0% | 0% | **89%** |
| LaBSE-After | 36% | 53% | 4% | **93%** | 1% | 3% | 2% | 1% | 0% | **7%** |

used heuristic, or have varied the criteria or thresholds of the heuristics arbitrarily from one study to another. Further, their work does not explore the potential of combining

multiple heuristics.

## 6.3 Related Work

Parallel data mined from the web at scale is often considered an alternative to human-created data in training Neural Machine Translation (NMT) models Costa-jussà et al. (2022); Bañón et al. (2020). CCAligned El-Kishky et al. (2020), CCMatrix Schwenk et al. (2021b) and ParaCrawl Bañón et al. (2020) are examples of such web-mined corpora, which cover Low-Resource Languages (LRLs) as well. However, quality audits of these corpora, as shown in Table 6.3, reveal substantial noise, which negatively impacts the NMT models Khayrallah and Koehn (2018).

Table 6.3: Noise types in parallel corpora, as identified by Khayrallah and Koehn (2018) (**A**), Bane et al. (2022) (**B**), Herold et al. (2022) (**C**), Kreutzer et al. (2022b) (**D**) and Ranathunga et al. (2024a) (**E**).

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

### 6.3.1 MultiPLMs for PDC

While employing a multiPLM for PDC is common, past research experimented with only one multiPLM at a time. For example, in the WMT2023 shared task, LASER2 was utilized to set the task baseline, whereas Steingrímsson (2023) and Gala et al. (2023) only used LaBSE. Therefore, the disparities across multiPLMs and biases specific to each multiPLM have not come into light. On the other hand, studies conducted by Ranathunga et al. (2024a) and Moon et al. (2023) reveal that using different multiPLMs for scoring and ranking parallel corpora, and training NMT models with the top-ranked corpora, results in a disparity. Moon et al. (2023) observe that

this is due to biases in multiPLMs, which tend to rank noisy parallel sentences highly. However, there has been no study to identify these biases.

### 6.3.2 Identifying Noise in Web-mined Corpora

Recent research used categorical labels to annotate translation pairs, aiming at quantifying the noise types in web-mined corpora. Kreutzer et al. (2022b) used their taxonomy to conduct manual audits on random samples from three web-mined datasets and reported substantial noise, specifically for LRLs. Ranathunga et al. (2024a)'s taxonomy is an extension of Kreutzer et al. (2022b)'s taxonomy. They first ranked the datasets based on embeddings from a multiPLM, and then selected random samples from the top and bottom portions. Their human evaluation based on the taxonomy reported that the quality of the parallel sentences varies heavily depending on the selected portion.

### 6.3.3 Heuristic-based PDC

Commonly used rule-based heuristics can be categorised into four groups as described below:

**De-duplication based *(Dedup)*:** Removing identical duplicates in the monolingual sides is a common practice (Costa-jussà et al., 2022), while De-duplicating after removing non-alpha characters and punctuations (Bala Das et al., 2023) could be found as its variants.

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

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

**Ratio-based :** Ratio-based heuristics identify and remove sentences that show significant structural imbalances between the source and target sentences. It is based on the assumption that well-aligned sentence pairs tend to maintain consistent ratios in terms of character count, word count, or token distribution. We observe three common types of ratio-based heuristics: (1) source-to-target sentence length ratio (*STRatio*) (Rossenbach et al., 2018; Gale and Church, 1993), (2) alpha-only words to sentence words ratio (*sentWRatio*) (Aulamo et al., 2020) and (3) alpha-only character ratio with respect to the sentence characters (*sentCRatio*) (Hangya and Fraser, 2018).

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

## 6.4 Methodology

We describe the approach taken in conducting the research work in this section. First, we describe the improvements proposed to the existing taxonomies (Section 6.4.1), secondly, the rationale for selecting the heuristics (Section 6.4.2) and finally, the methodology for conducting the human evaluation (Section 6.4.3).

### 6.4.1 Improved Taxonomy for Noise

Although translation pairs can have overlapping URLs, acronyms, etc, excessive inclusions of such content in a sentence (e.g. consider the sentence 'Contact: Diane Anderson 076-8268914, info@sandnasbadenscamping.se' (More examples are in Table C.1 in Appendix C) do not provide meaningful content for an NMT system to learn from. However, under Ranathunga et al. (2024a)'s taxonomy, such sentence pairs would likely be categorized as correct translations. Therefore, we define a new noise category *CCN* to capture such sentence pairs.

Secondly, we consider the upper limit for short sentences as five words, based on common observations. Finally, we improve the definition of *WL* (wrong language) to consider a threshold in determining whether a sentence pair should be marked as wrong language. Table 6.4 shows these changes. The complete list of these noise categories and example parallel sentences is available in Table C.2 and Table 6.1 respectively.

Table 6.4: A comparison of the improved taxonomy against Ranathunga et al. (2024a)'s. (only showing the changes)

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

### 6.4.2 Selection of Heuristics

Table 6.5 shows how the heuristics discussed in Section 6.3.3 may help in removing different noise categories. Note that de-duplication based heuristic cannot be associated with any noise category, as it does not apply to individual sentence pairs. In addition to de-duplication strategies discussed in Section 6.3.3, we introduce an n-gram based de-duplication, meaning that sentences would be removed if they overlap in a consecutive n-gram sequence.

### 6.4.3 Human Evaluation

We conduct a human evaluation to quantify the noise before and after applying the heuristics.

Table 6.5: Mapping between the noise category vs the noise mitigating heuristic.

| Noise Category | Short Label | Rule-based Heuristic |
|---|---|---|
| Not a language | NL | LID, sentWRatio, sentCRatio |
| Wrong language | WL | LID |
| Untranslated | UN | LID |
| Short Sentences | CS | sLength |
| Overlapping numbers/acronyms/URLs/email | CCN | LID, sentWRatio, sentCRatio |
| Wrong translation | X | STRatio (With a length difference) |
| Biolerplate translation | CB | STRatio |

**Selection of the Annotators:** We select annotators for this task who have translation or machine translation-related experience for a minimum of 2 years. Their first language is either `Si` or `Ta` and are from an educational background. Hence, we believe the annotators are competent for this task. Table C.3 in Appendix C shows the years of experience and the qualifications of those annotators who conducted this task.
**Resources Provided and Training:** All the annotators have had prior experience with a similar task. However, for this annotation work, we provide them with the definitions of the noise categories (Table C.2 in Appendix C) along with example sentence pairs (Table 6.1 in Appendix C) and the guideline in terms of a flowchart (Figure C.1 in Appendix C). First, we asked them to do a sample of 30 sentences as a training on the task, and review it. Then, the 1200 sentence pairs to be annotated were shared with each annotator via Google Sheets to be completed.
**Compensation:** They were paid for each sentence pair they annotated, which was the standard rate for such work.
**Annotation:** From the top 1000 samples in each of the ranked corpora using LASER3, XLM-R and LaBSE, we randomly select 100 parallel sentences for each language pair. We ask each translator to select an annotation category for each sentence pair using the error taxonomy discussed in Section 6.4.1. Each sentence pair was annotated by two translators to reduce potential bias caused by the individual translators. Inter-annotator agreement was calculated using the Pearson correlation coefficient, which resulted in En-Si 0.88, En-Ta 0.68 and Si-Ta 0.49.

## 6.5 Experiments

In this section, we describe the web-mined parallel datasets used (Section 6.5.1), rationale for selecting the multiPLMs (Section 6.5.2) and the experiments conducted (Section 6.5.3) during the research work.

### 6.5.1 Dataset

We conduct this research using the same three language pairs, En-Si, En-Ta, and Si-Ta. As a web-mined parallel corpus, we select CCMatrix (Artetxe and Schwenk, 2019a) and CCAligned (El-Kishky et al., 2020) as the web-mined corpora. Both these corpora include parallel data for the language pairs considered in the search. We describe the justification for the language selection in Section 2.6.

For the NMT experiments, we use the *dev* and *devtest* subsets from the Flores+ dataset[1] as validation and evaluation sets, respectively. Dataset statistics are provided in Table 6.6.

Table 6.6: Corpus statistics.

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

### 6.5.2 Selection of muliPLMs

We select LASER3, XLM-R, and LaBSE for obtaining embeddings for the sentences to determine the semantic similarity. XLM-R, different to others, was trained purely on monolingual data, but has proven to be useful for cross-lingual tasks as well Choi et al. (2021); Conneau et al. (2020a). All three models include En, Si, and Ta. Details of the multiPLMs XLM-R and LaBSE are in Section 4.3.2 in Chapter 4. For LASER3, the details as follows:

**LASER3** Heffernan et al. (2022) (L=12, H=1024, A=4, 250M)[2] is a multiPLM favourable for bitext mining and cross-lingual tasks. It improves over previous LASER2 versions by supporting more languages and enhancing alignment quality, but it still faces challenges in low-resource settings.

### 6.5.3 Heuristic-based PDC Experiments

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

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

[2] No of Layers, Hidden Layer Dimensions and No of Attention Heads are defined by L, H and A respectively.

**Deduplication**: We consider different granularities of de-duplication. ie. identical de-duplication *(dedup)*, de-duplicate by removing numbers only (*nums*) and removing both numbers and punctuations (*punctsNums*). Subsequently, we de-duplicate considering different n-gram spans, i.e. 4-grams, 5-grams, 6-grams and 7-grams.
**Length-based:** We filter short sentences less than five words[3]. While some research has suggested removing extremely long sentences (Minh-Cong et al., 2023a; Gala et al., 2023), we found that the percentage of longer sentences is lower and that removing them has a negligible effect. Thus, this result is not reported.
**LID-based:** We use a public LID model[4] to predict the language of each sentence. The predicted label is then used as a standalone heuristic (*LID*) and in combination with its associated prediction probability (*LIDThresh*), with threshold of 0.7[5].
**Ratio-based:** For *STRatio*, 0.79-1.39, 0.87-1.62 and 0.85-1.57 were selected as thresholds for En-Si, En-Ta and Si-Ta respectively. These were determined by calculating the mean and the standard deviation obtained from the validation set in the human-crafted En-Si-Ta trilingual dataset (Fernando et al., 2020). Following observations of Hangya and Fraser (2018), 0.6 was selected as the threshold for *sentWRatio* and *sentCRatio*.

### 6.5.4 NMT Experiments

First, a Sentencepiece[6] tokenizer with a vocabulary size of 25000 is trained. Then we use the fairseq toolkit (Ott et al., 2019) to model and train the transformer-based Seq-to-Seq NMT model until convergence. Hyperparameters used in the NMT experiments are shown in Table 6.7. The baseline NMT models are trained on the top 100,000 sentence pairs from the ranked corpus. We use ChrF++ Popović (2017) to report NMT results. We select ChrF++, as it evaluates the NMT output considering character n-grams, which is favourable for morphologically rich languages, compared to the conventional BLEU Papineni et al. (2002) metric, which evaluates the NMT output at the word-level.

## 6.6 Experimental Results

We train NMT models in the forward direction, taking `En` as the source in `En-Si` and `En-Ta` and `Si` as the source in case of the `Si-Ta` language-pair. The results of the experiments are in Table 6.8. As evident from these tables, as well as from Figure 6.1, NMT results across different multiPLMs show a great disparity in the baseline NMT scores for `En-Si` and `En-Ta` language pairs. In the following sub-sections, we discuss how the use of heuristics is useful in mitigating this disparity and improving overall NMT results.

[3]This is the commonly used threshold in the existing work.
[4]https://github.com/facebookresearch/fairseq/tree/nllb
[5]Thresholds below 0.7 reduce NMT results
[6]https://github.com/google/sentencepiece

Table 6.7: Training parameters for NMT experiments.

| **Hyperparameter** | **Argument value** |
|---|---|
| encoder/decoder Layers | 6 |
| encoder/decoder attention heads | 4 |
| encoder-embed-dim | 512 |
| decoder-embed-dim | 512 |
| encoder-ffn-embed-dim | 2048 |
| decoder-ffn-embed-dim | 2048 |
| dropout | 0.4 |
| attention-dropout | 0.2 |
| optimizer | adam |
| Adam $\beta_1$, Adam $\beta_2$ | 0.9, 0.99 |
| warmup-updates | 4000 |
| warmup-init-lr | 1e-7 |
| learning rate | 1e-3 |
| batch-size | 32 |
| patience | 6 |
| fp16 | True |

Table 6.8: NMT results obtained after applying heuristics in isolation and in combination in the ablation study. The values in bold indicate the highest NMT score obtained for a given heuristic class or from the heuristic combination. The values underlined are the highest among the individual heuristics. Highlighted in green are the overall best values. Here **DD+PN** is *Deduplication+punctNums*, **SL** is *sLength* and **LT** is *LIDThresh*. Here NA would be when the particular experiment is not applicable for that language pair or the dataset.

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

### 6.6.1 Impact of Heuristics on NMT Results

In this section, we take the individual heuristic at a time and discuss its impact on the final NMT performance. The results are analysed with respect to several dimensions, including the type of heuristic applied (S/T/ST)[7], the language pairs considered, and the datasets employed, as applicable.

#### 6.6.1.1 Impact of De-duplication based PDC

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

We apply our newly introduced ngram-based de-duplication *dedup+ngram*, on top of *dedup*. We find that for each result column, there is a *dedup+ngram* result that outperforms the best results obtained with the corresponding *dedup* result. To observe the impact of *n* value on the NMT result, we plot Figure 6.2 and Figure 6.3 showing the percentage of *ngram* experiments exceeding the highest *dedup*, with respect to the multiPLM and language-pair respectively. We observe a consistent pattern - n=5 or 6 perform the best in a majority of cases. We believe that 4-gram results in an overly aggressive de-duplication. However, the exact ngram depends on the corpus characteristics.

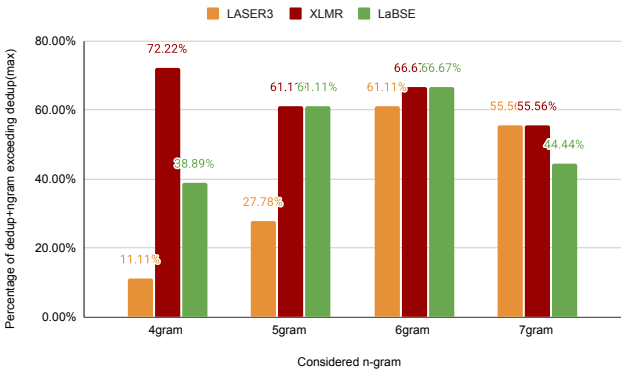


Figure 6.2: Percentage of *dedup+ngram* experiments exceeding the best result of *dedup* for each *multiPLM*

We observe that *dedup+punctNums* outperforms *dedup+nums* and *dedup* in 78% and 89% of the experiments (respectively), proving that *(dedup+punctNums)* to be more impactful. Finally, we analyse the impact of *dedup+puntsNums+ngram*[8]. Compared to

[7] S – Source side, T – Target side, and ST – Source and Target sides

[8] *puntsNums* and *ngram* has been applied on top of *dedup*.

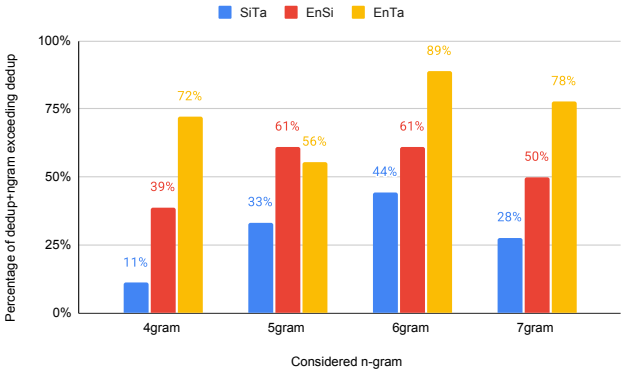


Figure 6.3: Percentage of *dedup+ngram* experiments exceeding the best result of *dedup* with respect to the Language-pair.

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

#### 6.6.1.2 Impact of Length-Based PDC

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

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

#### 6.6.1.3 Impact of LID-Based PDC

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

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

[9] *dedup*, *dedup+ngram*, *nums*, and *dedup+punctNums*

#### 6.6.1.4 Impact of Ratio-based PDC

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

### 6.6.2 Summary of Heuristic-based PDC

Here, we analyze our results and present conclusions regarding the most impactful individual heuristic, as well as the heuristic combinations that yield the best overall performance. Finally, we discuss the impact of the heuristics on the disparity among the NMT models, trained using the ranked parallel corpora.

# CHAPTER 7

# DISCUSSION

In this thesis, we investigated Data Augmentation (DA) techniques aimed at addressing the data scarcity issue in low-resource Neural Machine Translation (NMT). We primarily considered three principal DA strategies: word or phrase replacement-based augmentation, bitext mining and Parallel Data Curation (PDC) and implemented techniques to enhance the overall reliability of the resulting parallel corpora. At the beginning of this thesis, we defined four research objectives centred on this.

## 7.1 Research Objectives

Our thesis aimed at addressing four main research objectives, which has been described in Chapter 1. In this section, we will revisit these objectives and reflect on how we addressed them:

**RO1. Propose and implement an algorithm to generate synthetic parallel sentences to augment out-of-vocabulary terms.**

We addressed this in Chapter 3. Our DA approach was aimed at generating synthetic parallel sentences by augmenting Out-of-Vocabulary (OOV) terms. OOV terms are twofold: rare words and unseen words in the training corpus. Rare words are words which exist in the training corpus but with a low frequency. Unseen words are vocabulary that do not exist at all in the training corpus. Since the Neural Machine Translation (NMT) systems do not encounter these words during the training time at all or in an adequate frequency, during inference, they fail to produce reliable translations for the sequences containing such words.

In our approach, we augmented both types of OOV terms by substituting them in existing parallel sentences to provide novel contexts, conforming to syntactic and semantic constraints. As syntactic constraints, we considered POS and morphological agreement and as semantic constraints, we considered word and sentence semantic similarity. Further, we optimized the word embeddings and determined the semantic similarity by means of a linear transformation in a post-processing step. Finally, we incorporated forward and backward trained Language Models (LMs) to validate the replacement context. Thereby, we ensured that *high-quality* synthetic parallel sentences were induced, that were plausible in the respective languages. We empirically showed that our augmentation approach not only produced reliable translations for the sequences containing OOV terms but also improved the overall NMT results for the language pair.

Our approach benefits the low-resource NMT in two ways. Firstly, it addresses the data scarcity problem by inducing synthetic parallel sentences which is rich in lexical

diversity. Thereby the overall NMT score for the language pair is improved. Secondly, we propose that this method is a viable approach for Low-Resource Languages (LRLs) and empirically prove that even with using semantic constraints, comparable results can be obtained.

**RO2: Conduct an empirical Study to determine the impact of different characteristics of the Pre-trained Multilingual Language Models on the Document Alignment and Sentence Alignment tasks for LRLs.**

We addressed this research objective in Chapter 4. Although bitext-mining from web data is a viable DA technique to address the data scarcity problem, for LRLs, the produced parallel data sets are often noisy. This is due to the suboptimal performance of the two most crucial subtasks in the bitext-mining pipeline, which are document alignment and sentence alignment. The recent approach for determining the alignment between the parallel documents or sentences is by considering the semantic similarity calculated between the embeddings obtained from a multiPLM for the source side and target side documents or sentences. We conducted an empirical study and evaluated the effectiveness of the type of multiPLMs favourable to obtain embeddings for the document and sentence alignment tasks. Secondly, we investigated whether these alignment tasks could be further improved with a weighting mechanism derived from small-scale bilingual lexicons, which improved the scoring function deciding the alignment among the documents or sentences.

Our results showed that embeddings obtained from a multiPLM, which had undergone fine-tuning with parallel data, led to producing the best gains for document alignment and sentence alignment tasks. Additionally, we observed that dictionary-based scoring function improvements are insignificant in such a setting. Hence we conclude that multiPLM trained explicitly with a cross-lingual objective is favourable to improve the performance for LRLs.

**RO3. Improve the cross-lingual representations of existing multiPLMs to obtain High-Quality parallel sentences from the parallel sentence alignment task.**

We addressed this research objective in Chapter 5, taking the motivation from the findings in RO2. Although multiPLMs trained using Masked Language Modelling (MLM) objective improved cross-lingual performance for several downstream Natural Language Processing (NLP) tasks, their success in sentence-retrieval tasks is suboptimal. The Translation Language Modelling (TLM) was proposed in a continual pre-training step to optimize these MLM-trained embeddings for cross-lingual tasks. However, during masking, both MLM and TLM selected tokens for masking randomly, disregarding their linguistic prominence in the sequence. In our approach, we introduced a novel masking strategy, *Linguistic Entity Masking* (LEM), to be used in a 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. We evaluated the effectiveness of LEM using two evaluation tasks, namely sentence alignment and parallel data curation in a LRL setting. Empirically, we proved that LEM improved the existing XLM-R representations, using only a 50k parallel dataset size.

**RO4. Exploring Parallel Data Curation (PDC) techniques to extract high-quality parallel sentences from web-mined parallel corpora.**

We addressed this research objective in Chapter 6. Parallel Data Curation (PDC) aims at filtering *high-quality* parallel sentences from existing web-mined corpora. The recent PDC techniques follow a scoring and ranking mechanism using embeddings obtained from a multiPLM. The NMT systems are then trained with the top-ranked parallel sentences. In line with the existing research, we observed that using embeddings obtained from different multiPLMs for scoring results in a notable disparity in the performance of the NMT systems. This discrepancy stems from inherent biases in multiPLMs, producing embeddings that result in higher semantic similarity scores even for noisy parallel sentence pairs.

We conducted heuristic-based PDC, and with each heuristic and heuristic combinations, we analysed the impact on the final NMT performance. Our ablation experiments showed that such a heuristic combination led to optimal NMT results and minimised the disparity. Our findings suggested that ranking with multiPLMs should be complemented with heuristic-based PDC to ensure more reliable and consistent NMT results. Furthermore, through human evaluation, we quantified the noise in the ranked corpus, before and after applying heuristics and observed that qualitatively, the curated parallel corpus maintained a comparable level of quality.

## 7.2 Future Work

We explored several DA strategies and proposed techniques for improving them. Nevertheless, in a LRL setting, we encountered limitations as discussed in detail under the respective chapters. In this section, we discuss the possible paths for expanding our research.

### 7.2.1 Inducing Synthetic Sentences

In the word or replacement-based augmentation work, to decide on the suitable context for the OOV to be replaced, we relied on syntactic and semantic features. However, to obtain `Sinhala` PoS tag information, we relied on a SVM model where the prediction was independent of the context of its usage. Therefore, we plan to improve the PoS tagger, by fine-tuning a multiPLM for the sequence labelling task to better optimise for accurate PoS tag predictions. Secondly, to determine the semantic similarity, we relied on FastText embeddings. This can be improved by training a BERT-based model for

Sinhala to obtain contextualised embeddings which are richer in terms of semantic representations than the static word embeddings. The two improvements which we take up as future work will not only benefit the current DA work but also serve other Sinhala related NLP downstream tasks as well.

### 7.2.2 Effectiveness of multiPLMs on Document Alignment and Sentence Alignment tasks

In this empirical study, we evaluated the effectiveness of embeddings returned by encoder-based multiPLMs for the document alignment and sentence alignment tasks for LRL pairs. Several studies have been conducted to utilize embeddings extracted from sequence-to-sequence models (Ni et al., 2022) for cross-lingual tasks. Complementing this work, decoder-based generative Large Language Models (LLMs) have been explored in similar information retrieval ranking tasks (Sun et al., 2025) as well. Taking the motivation from these research directions, we will extend our empirical study to explore the suitability of utilising sequence-to-sequence and decoder-based LLMs for the document alignment and sentence alignment tasks for LRLs.

### 7.2.3 Improving Cross-Lingual Representations

In this work, we proposed the Linguistic Entity Masking (LEM) strategy and empirically demonstrated its effectiveness in improving the cross-lingual representations of existing multiPLMs. Our approach improves the multiPLM for each language pair separately by applying the LEM strategy. As future work, we will investigate whether a unified multilingual model can be continually pre-trained using LEM strategy across multiple LRLs, to improve cross-lingual representations across those languages.

We also foresee that ongoing advancements in NLP will yield more accurate Named Entity Recognition (NER) and Part-of-Speech (POS) tagging tools for LRLs. Upon the availability of such tools, we intend to re-evaluate the performance of LEM under these improved linguistic tools, assessing its robustness. In the current study, the considered languages belong to three diverse language families. We aim to conduct an ablation study to evaluate the effectiveness of LEM strategy across several LRLs belonging to different language families and to assess its potential in capturing broader linguistic variation. In addition, provided that we acquire computation resources, we can explore the effectiveness of LEM in the pre-training phase of a multilingual model as well.

### 7.2.4 Parallel Data Curation

We observed that the heuristic-based PDC effectively reduced the disparity in NMT scores, eliminating the multiPLM biased noise. However, from the human evaluation, it was revealed that still noisy parallel sentences in the form of untranslated text still

exist after the filtration. As part of future work, we intend to investigate the use of classification-based approaches to identify and eliminate such residual noise. While classifiers have shown promise in PDC tasks, their effectiveness in LRL settings remains suboptimal. We therefore aim to enhance classification-based techniques to further improve the quality of curated parallel datasets for LRLs.

Further, it was observed that the combined-heuristic filtering step removed approximately 60%–70% of the parallel sentence pairs. Motivated by the findings of Steingrímsson et al. (2023), we plan to examine whether this discarded subset can still contribute positively to low-resource NMT systems, potentially through alternative strategies such as data selection, confidence-weighted training, or semi-supervised learning.

## 7.3 Chapter Summary

In this chapter, we provide a comprehensive reflection on how the research objectives outlined at the outset of this thesis have been systematically addressed through the conducted studies. We revisit each objective and highlight how the proposed methodologies contributed towards advancing the considered DA techniques. The chapter outlines the key findings and empirical insights derived from the experimental results.

Furthermore, we critically examine the limitations encountered across the various components of this research and propose several directions for future work. The proposed future work paves the way for potential extensions that could build upon the current study.

# CHAPTER 8

# CONCLUSION

Neural Machine Translation (NMT) systems for Low-Resource Languages (LRLs) continue to underperform compared to their High-Resource Language (HRL) counterparts, even for the state-of-the-art architectures. This can be primarily attributed to two sub-problems: (1) the parallel data scarcity problem and (2) the presence of noise in the existing parallel data.

We explored several existing Data Augmentation (DA) strategies, and recommended improvements, favourable for LRLs to yield *high-quality* parallel data in return to improve NMT scores. First, we improved word or phrase replacement-based augmentation for generating synthetic parallel sentences by imposing both syntactic and semantic constraints. Thereby, we ensured that the augmented sentences were grammatically well-formed and semantically correct in the respective languages. Next, we conducted an ablation study to determine the type of multiPLMs favourable to be used in the document alignment and sentence alignment tasks in the bitext mining pipeline. Thereby, we recommended that a multiPLM undergoing a fine-tuning step with parallel data was favourable to produce optimal results in the considered tasks to eventually return *high-quality* parallel sentences. Subsequently, we improved cross-lingual embeddings returned by an existing multiPLM with Linguistic Entity Masking (LEM) strategy. This improved the sentence alignment performance in the bitext mining pipeline. Final work was aimed towards curating existing web-mined noisy parallel corpora. We empirically proved that scoring and ranking using embeddings obtained from a multiPLM should be combined with huristic-based PDC to obtain a *high-quality* parallel sentences and to achieve optimal NMT gains.

Extensive empirical evaluations demonstrated that each of these approaches contributed to the generation of *high-quality* parallel sentence pairs and led to measurable improvements in NMT performance across the three LRL pairs. The artefacts produced during the course of this research—including source code, curated parallel datasets, and cross-lingual improved encoder models were made publicly available to benefit the broader research community. Furthermore, the core findings and contributions were disseminated through publications in peer-reviewed journals and international conferences.

# REFERENCES


Abdul-Rauf, S. and Schwenk, H. (2009). On the use of comparable corpora to improve smt performance. In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pages 16–23.

Abdulmumin, I., Galadanci, B. S., and Isa, A. (2020). Enhanced back-translation for low resource neural machine translation using self-training. In International Conference on Information and Communication Technology and Applications, pages 355–371. Springer.

Açarçiçek, H., Çolakoğlu, T., Hatipoğlu, P. E. A., Huang, C. H., and Peng, W. (2020). Filtering noisy parallel corpus using transformers with proxy task learning. In Proceedings of the Fifth Conference on Machine Translation, pages 940–946.

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In COLING 2018, 27th International Conference on Computational Linguistics, pages 1638–1649.

Alam, M. M. I., Ahmadi, S., and Anastasopoulos, A. (2024). A morphologically-aware dictionary-based data augmentation technique for machine translation of under-represented languages. arXiv preprint arXiv:2402.01939.

Allen B. Tucker, J. and Nirenburg, S. (1984). Machine translation: A contemporary view. Annual Review of Information Science and Technology, 19:129.

Aoyama, T. and 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, pages 195–201.

Araabi, A. and Monz, C. (2020). Optimizing transformer for low-resource neural machine translation. In Proceedings of the 28th International Conference on Computational Linguistics, pages 3429–3435.

Artetxe, M., Labaka, G., and Agirre, E. (2018a). Unsupervised statistical machine translation. In ACL.

Artetxe, M., Labaka, G., Lopez-Gazpio, I., and Agirre, E. (2018b). Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation. arXiv preprint arXiv:1809.02094.

Artetxe, M. and Schwenk, H. (2019a). Margin-based parallel corpus mining with multilingual sentence embeddings. In Proceedings of the 57th Annual Meeting of

the Association for Computational Linguistics, pages 3197–3203. Association for Computational Linguistics.

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

Aulamo, M., De Gibert, O., Virpioja, S., and Tiedemann, J. (2023). Unsupervised feature selection for effective parallel corpus filtering. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation, pages 31–38.

Aulamo, M., Virpioja, S., and Tiedemann, J. (2020). Opusfilter: A configurable parallel corpus filtering toolbox. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 150–156.

Azpeitia, A., Etchegoyhen, T., and Garcia, E. M. (2017). Weighted set-theoretic alignment of comparable sentences. In Proceedings of the 10th Workshop on Building and Using Comparable Corpora, pages 41–45.

Azpeitia, A., Etchegoyhen, T., and Garcia, E. M. (2018). Extracting parallel sentences from comparable corpora with stacc variants. In Proceedings of the 11th Workshop on Building and Using Comparable Corpora, pages 48–52.

Bahdanau, D., Cho, K. H., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In 3rd International Conference on Learning Representations, ICLR 2015.

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

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

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., et al. (2020). Paracrawl: Web-scale acquisition of parallel corpora. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4555–4567.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606.

Bouamor, H. and Sajjad, H. (2018). H2@ bucc18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In Proc. Workshop on Building and Using Comparable Corpora, pages 43–47.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. Computational linguistics, 19(2):263–311.

Brown, P. F., Lai, J. C., and Mercer, R. L. (1991). Aligning sentences in parallel corpora. In 29th Annual Meeting of the Association for Computational Linguistics, pages 169–176, Berkeley, California, USA. Association for Computational Linguistics.

Buck, C. and Koehn, P. (2016a). Findings of the WMT 2016 bilingual document alignment shared task. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 554–563, Berlin, Germany. Association for Computational Linguistics.

Buck, C. and Koehn, P. (2016b). Quick and reliable document alignment via tf/idf-weighted cosine distance. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 672–678.

Burchell, L., de Gibert, O., Arefyev, N., Aulamo, M., Bañón, M., Fedorova, M., Guillou, L., Haddow, B., Hajič, J., Henriksson, E., et al. (2025). An expanded massive multilingual dataset for high-performance language technologies. arXiv preprint arXiv:2503.10267.

Carlson, L. and Vilkuna, M. (1990). Independent transfer using graph unification. In COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics.

Caswell, I., Chelba, C., and Grangier, D. (2019). Tagged back-translation. In Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers), page 53. Association for Computational Linguistics.

Chaudhary, V., Tang, Y., Guzmán, F., Schwenk, H., and Koehn, P. (2019). Low-resource corpus filtering using multilingual sentence embeddings. WMT 2019, page 261.

Chen, J. and Nie, J.-Y. (2000). Parallel web text mining for cross-language ir. In Content-Based Multimedia Information Access-Volume 1, pages 62–77. RIAO.

Chen, J., Tam, D., Raffel, C., Bansal, M., and Yang, D. (2023). An empirical survey of data augmentation for limited data learning in nlp. Transactions of the Association for Computational Linguistics, 11:191–211.

Cho, K., van Merriënboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724–1734.

Choi, H., Kim, J., Joe, S., Min, S., and Gwon, Y. (2021). Analyzing zero-shot cross-lingual transfer in supervised nlp tasks. In 2020 25th International Conference on Pattern Recognition (ICPR), pages 9608–9613. IEEE.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020a). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020b). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. Advances in neural information processing systems, 32.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al. (2022). No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.

Dabre, R., Chu, C., and Kunchukuttan, A. (2020). A survey of multilingual neural machine translation. ACM Computing Surveys (CSUR), 53(5):1–38.

Dabre, R., Fujita, A., and Chu, C. (2019). Exploiting multilingualism through multi-stage fine-tuning for low-resource neural machine translation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1410–1416.

Dara, A. A. and Lin, Y.-C. (2016). Yoda system for wmt16 shared task: Bilingual document alignment. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 679–684.

De Gibert, O., Nail, G., Arefyev, N., Bañón, M., Van Der Linde, J., Ji, S., Zaragoza-Bernabeu, J., Aulamo, M., Ramírez-Sánchez, G., Kutuzov, A., et al. (2024). A new massive multilingual dataset for high-performance language technologies. In Proceedings of the 2024 Joint International Conference on Computational

Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 1116–1128.

de Silva, N. (2019). Survey on publicly available sinhala natural language processing tools and research. arXiv preprint arXiv:1906.02358.

de Silva, N. (2023). Survey on Publicly Available Sinhala Natural Language Processing Tools and Research. arXiv preprint arXiv:1906.02358v20.

de Silva, N. (2025). Survey on Publicly Available Sinhala Natural Language Processing Tools and Research. arXiv preprint arXiv:1906.02358v24.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019a). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019b). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dhananjaya, V., Demotte, P., Ranathunga, S., and 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, pages 7377–7385.

Dhar, P., Bisazza, A., and van Noord, G. (2021). Optimal word segmentation for neural machine translation into dravidian languages. In Proceedings of the 8th Workshop on Asian Translation (WAT2021), pages 181–190.

Duan, S., Zhao, H., Zhang, D., and Wang, R. (2020). Syntax-aware data augmentation for neural machine translation. arXiv preprint arXiv:2004.14200.

El-Kishky, A., Chaudhary, V., Guzmán, F., and 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), pages 5960–5969.

El-Kishky, A. and Guzmán, F. (2020). Massively multilingual document alignment with cross-lingual sentence-mover's distance. In Proceedings of the 1st Conference of the

Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 616–625, Suzhou, China. Association for Computational Linguistics.

Epaliyana, K., Ranathunga, S., and Jayasena, S. (2021). Improving back-translation with iterative filtering and data selection for sinhala-english nmt. In 2021 Moratuwa Engineering Research Conference (MERCon), pages 438–443. IEEE.

Espla-Gomis, M., Forcada, M. L., Ortiz-Rojas, S., and Ferrández-Tordera, J. (2016). Bitextor's participation in wmt'16: shared task on document alignment. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 685–691.

Etchegoyhen, T. and Gete, H. (2020). Handle with care: A case study in comparable corpora exploitation for neural machine translation. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 3799–3807.

Fadaee, M., Bisazza, A., and Monz, C. (2017). Data augmentation for low-resource neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 567–573.

Fadaee, M. and Monz, C. (2018). Back-translation sampling by targeting difficult words in neural machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 436–446.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., and Joulin, A. (2020). Beyond english-centric multilingual machine translation.

Farhath, F., Ranathunga, S., Jayasena, S., and Dias, G. (2018a). Integration of bilingual lists for domain-specific statistical machine translation for sinhala-tamil. In 2018 Moratuwa Engineering Research Conference (MERCon), pages 538–543. IEEE.

Farhath, F., Theivendiram, P., Ranathunga, S., Jayasena, S., and Dias, G. (2018b). Improving domain-specific smt for low-resourced languages using data from different domains. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2020). Language-agnostic bert sentence embedding. arXiv preprint arXiv:2007.01852.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and 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), pages 878–891.

Fernando, A. and Dias, G. (2021). Building a linguistic resource: A word frequency list for sinhala. In Proceedings of the 18th International Conference on Natural Language Processing (ICON), pages 606–610.

Fernando, A. and Ranathunga, S. (2021). Data augmentation to address out of vocabularyproblem in low resource sinhala english neural machine translation. In Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation, pages 61–70.

Fernando, A. and Ranathunga, S. (2025). Linguistic entity masking to improve cross-lingual representation of multilingual language models for low-resource languages. Knowledge and Information Systems.

Fernando, A., Ranathunga, S., and de Silva, N. (2025). Improving the quality of web-mined parallel corpora of low-resource languages using debiasing heuristics. arXiv preprint arXiv:2502.19074.

Fernando, A., Ranathunga, S., and Dias, G. (2020). Data augmentation and terminology integration for domain-specific sinhala-english-tamil statistical machine translation. arXiv preprint arXiv:2011.02821.

Fernando, A., Ranathunga, S., Sachintha, D., Piyarathna, L., and Rajitha, C. (2023). Exploiting bilingual lexicons to improve multilingual embedding-based document and sentence alignment for low-resource languages. Knowledge and Information Systems, 65(2):571–612.

Fernando, S. and Ranathunga, S. (2018). Evaluation of different classifiers for sinhala pos tagging. In 2018 Moratuwa Engineering Research Conference (MERCon), pages 96–101. IEEE.

Fernando, S., Ranathunga, S., Jayasena, S., and 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), pages 173–182.

Fonseka, T., Naranpanawa, R., Perera, R., and Thayasivam, U. (2020). English to sinhala neural machine translation. In 2020 International Conference on Asian Language Processing (IALP), pages 305–309. IEEE.

Fung, P. and Cheung, P. (2004). Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and e. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 57–63.

Gala, J., Chitale, P. A., AK, R., Gumma, V., Doddapaneni, S., Kumar, A., Nawale, J., Sujatha, A., Puduppully, R., Raghavan, V., et al. (2023). Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. arXiv preprint arXiv:2305.16307.

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

Gao, Y., Hou, F., Jahnke, H., and Wang, R. (2023). Data augmentation with diversified rephrasing for low-resource neural machine translation. In Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 35–47.

Garcia, X., Niu, Y., and Specia, L. (2023). Low-resource domain-robust unsupervised machine translation via multi-phase adaptation. In Findings of ACL.

Germann, U. (2016). Bilingual document alignment with latent semantic indexing. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 692–696, Berlin, Germany. Association for Computational Linguistics.

Golchin, S., Surdeanu, M., Tavabi, N., and 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), pages 13–21.

Gomes, L. and Lopes, G. (2016). First steps towards coverage-based document alignment. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 697–702.

Gowda, T., Zhang, Z., Mattmann, C., and May, J. (2021). Many-to-English machine translation tools, data, and pretrained models. In Ji, H., Park, J. C., and Xia, R., editors, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, pages 306–316, Online. Association for Computational Linguistics.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., and Fan, A. (2022). The flores-101 evaluation benchmark for low-resource and multilingual machine translation. Transactions of the Association for Computational Linguistics, 10:522–538.

Grégoire, F. and Langlais, P. (2017). Bucc 2017 shared task: a first attempt toward a deep learning framework for identifying parallel sentences in comparable corpora. In Proceedings of the 10th Workshop on Building and Using Comparable Corpora, pages 46–50.

Guoa, M., Shenb, Q., Yanga, Y., Gea, H., Cera, D., Abregoa, G. H., Stevensa, K., Constanta, N., Sunga, Y.-H., Stropea, B., et al. (2018). Effective parallel corpus mining using bilingual sentence embeddings. WMT 2018, page 165.

Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V., and Ranzato, M. (2019). The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Ha, T.-L., Niehues, J., and Waibel, A. (2016). Toward multilingual neural machine translation with universal encoder and decoder. In Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., Cattoni, R., and Federico, M., editors, Proceedings of the 13th International Conference on Spoken Language Translation, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Haddow, B., Bawden, R., Miceli-Barone, A. V., Helcl, J., and Birch, A. (2022). Survey of low-resource machine translation. Computational Linguistics, 48(3):673–732.

Hangya, V. and Fraser, A. (2018). An unsupervised system for parallel corpus filtering. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pages 882–887.

Hangya, V. and Fraser, A. (2019). Unsupervised parallel sentence extraction with parallel segment detection helps machine translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1224–1234.

Heffernan, K., Çelebi, O., and Schwenk, H. (2022). Bitext mining using distilled sentence representations for low-resource languages. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 2101–2112.

Herold, C., Rosendahl, J., Vanvinckenroye, J., and Ney, H. (2022). Detecting various types of noise for neural machine translation. In Findings of the Association for Computational Linguistics: ACL 2022, pages 2542–2551.

Hu, J., Johnson, M., Firat, O., Siddhant, A., and Neubig, G. (2021a). 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, pages 3633–3643.

Hu, J., Johnson, M., Firat, O., Siddhant, A., and Neubig, G. (2021b). 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, pages 3633–3643.

Ion, R., Ceauşu, A., and Irimia, E. (2011). An expectation maximization algorithm for textual unit alignment. In Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, pages 128–135.

Isabelle, P. and Macklovitch, E. (1986). Transfer and mt modularity. In Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics.

Isuranga, U., Sandaruwan, J., Athukorala, U., and Dias, G. (2020). Improved cross-lingual document similarity measurement.

Iyyer, M., Manjunatha, V., Boyd-Graber, J., and Daumé III, H. (2015). Deep unordered composition rivals syntactic methods for text classification. In Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers), pages 1681–1691.

Jain, M., Punia, R., and Hooda, I. (2020). Neural machine translation for tamil to english. Journal of Statistics and Management Systems, 23(7):1251–1264.

Jakubina, L. and Langlais, P. (2016). Bad luc@ wmt 2016: a bilingual document alignment platform based on lucene. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 703–709.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5:339–351.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, 8:64–77.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1700–1709.

Khayrallah, H. and Koehn, P. (2018). On the impact of various types of noise on neural machine translation. In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pages 74–83.

Klein, G., Kim, Y., Deng, Y., Nguyen, V., Senellart, J., and Rush, A. M. (2018). Opennmt: Neural machine translation toolkit. In Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), pages 177–184.

Kocmi, T., Zouhar, V., Federmann, C., and Post, M. (2024). Navigating the metrics maze: Reconciling score magnitudes and accuracies. In Ku, L.-W., Martins, A., and Srikumar, V., editors, Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In Proceedings of machine translation summit x: papers, pages 79–86.

Koehn, P., Chaudhary, V., El-Kishky, A., Goyal, N., Chen, P.-J., and Guzmán, F. (2020). Findings of the wmt 2020 shared task on parallel corpus filtering and alignment. In Proceedings of the Fifth Conference on Machine Translation, pages 726–742.

Koehn, P., Guzmán, F., Chaudhary, V., and Pino, J. (2019). Findings of the wmt 2019 shared task on parallel corpus filtering for low-resource conditions. In Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), pages 54–72.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, pages 177–180.

Koehn, P., Khayrallah, H., Heafield, K., and Forcada, M. L. (2018). Findings of the wmt 2018 shared task on parallel corpus filtering. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pages 726–739.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation, pages 28–39.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pages 127–133.

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Suarez, P. O., Orife, I., Ogueji, K., Rubungo, A. N., Nguyen, T. Q., Müller, M., Müller, A., Muhammad, S. H., Muhammad, N., Mnyakeni, A., Mirzakhalov, J., Matangira, T., Leong, C., Lawson, N., Kudugunta, S., Jernite, Y., Jenny, M., Firat, O., Dossou, B. F. P., Dlamini, S., de Silva, N., Çabuk Ballı, S., Biderman, S., Battisti, A., Baruwa, A., Bapna, A., Baljekar, P., Azime, I. A., Awokoya, A., Ataman, D., Ahia, O., Ahia, O., Agrawal, S., and Adeyemi, M. (2022a). Quality at a glance: An audit of web-crawled multilingual datasets. Transactions of the Association for Computational Linguistics, 10:50–72.

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A. A., Subramani, N., Sokolov, A., Sikasote, C., et al. (2022b). Quality at a glance: An audit of web-crawled multilingual datasets. Transactions of the Association for Computational Linguistics, 10:50–72.

Krupakar, H. and Milton, R. S. (2016). Improving the performance of neural machine translation involving morphologically rich languages. ArXiv, abs/1612.02482.

Kudugunta, S., Caswell, I., Zhang, B., Garcia, X., Xin, D., Kusupati, A., Stella, R., Bapna, A., and Firat, O. (2024). Madlad-400: A multilingual and document-level large audited dataset. Advances in Neural Information Processing Systems, 36.

Kumarasinghe, K., Dias, G., and Herath, I. (2021). Sinmorphy: A morphological analyzer for the sinhala language. In 2021 Moratuwa Engineering Research Conference (MERCon), pages 681–686. IEEE.

Kvapilíková, I., Artetxe, M., Labaka, G., Agirre, E., and Bojar, O. (2020). Unsupervised multilingual sentence embeddings for parallel corpus mining. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 255–262.

Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E. (2017). Race: Large-scale reading comprehension dataset from examinations. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 785–794.

Lakmal, D., Ranathunga, S., Peramuna, S., and Herath, I. (2020). Word embedding evaluation for sinhala. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 1874–1881.

Lample, G. and Conneau, A. (2018). Phrase-based & neural unsupervised machine translation. In EMNLP.

Latief, A. D., Jarin, A., Yaniasih, Y., Afra, D. I. N., Nurfadhilah, E., Pebiana, S., Hidayati, N. N., and Fajri, R. (2024). Latest research in data augmentation for low resource language text translation: A review. In 2024 International Conference on Computer, Control, Informatics and its Applications (IC3INA), pages 185–190. IEEE.

Lee, E.-S. A., Thillainathan, S., Nayak, S., Ranathunga, S., Adelani, D. I., Su, R., and McCarthy, A. D. (2022). Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? arXiv preprint arXiv:2203.08850.

Leong, C., Wong, D. F., and Chao, L. S. (2018). Um-paligner: Neural network-based parallel sentence identification model. In 11th Workshop on Building and Using Comparable Corpora, page 53.

Leveling, J., Ganguly, D., Dandapat, S., and Jones, G. (2012). Approximate sentence retrieval for scalable and efficient example-based machine translation. In Kay, M. and Boitet, C., editors, Proceedings of COLING 2012, pages 1571–1586, Mumbai, India. The COLING 2012 Organizing Committee.

Levine, Y., Lenz, B., Lieber, O., Abend, O., Leyton-Brown, K., Tennenholtz, M., and Shoham, Y. (2020). Pmi-masking: Principled masking of correlated spans. In International Conference on Learning Representations.

Li, B. and Gaussier, E. (2013). Exploiting comparable corpora for lexicon extraction: Measuring and improving corpus quality. In Building and using comparable corpora, pages 131–149. Springer.

Liu, D., Ma, N., Yang, F., and Yang, X. (2019). A survey of low resource neural machine translation. In 2019 4th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), pages 39–393. IEEE.

Liu, H., Hou, R., and Lepage, Y. (2024). High-quality data augmentation for low-resource nmt: Combining a translation memory, a gan generator, and filtering. arXiv preprint arXiv:2408.12079.

Liu, X., He, J., Liu, M., Yin, Z., Yin, L., and Zheng, W. (2023). A scenario-generic neural machine translation data augmentation method. electronics 2023, 12, 2320. doi.org/10.3390/electronics12102320, 4.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 8:726–742.

Lopes, A., Farajian, M. A., Bawden, R., Zhang, M., and Martins, A. F. (2020). Document-level neural mt: A systematic comparison. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, pages 225–234.

Lu, H., Huang, H., Zhang, D., Wei, F., and Lam, W. (2024). Revamping multilingual agreement bidirectionally via switched back-translation for multilingual neural machine translation. In Findings of the Association for Computational Linguistics: EACL 2024, pages 264–275.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1412–1421.

Ma, X. (2006). Champollion: A robust parallel text sentence aligner. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), pages 489–492, Genoa, Italy. European Language Resources Association (ELRA).

Ma, X. and Liberman, M. (1999). Bits: A method for bilingual text search over the web. In Machine Translation Summit VII, pages 538–542.

Mager, M., Bhatnagar, R., Neubig, G., Vu, N. T., and Kann, K. (2023). Neural machine translation for the indigenous languages of the americas: An introduction. In Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 109–133.

Mahata, S., Das, D., and Bandyopadhyay, S. (2017). Bucc2017: A hybrid approach for identifying parallel sentences in comparable corpora. In Proceedings of the 10th Workshop on Building and Using Comparable Corpora, pages 56–59.

Maimaiti, M., Liu, Y., Luan, H., and Sun, M. (2022). Data augmentation for low-resource languages nmt guided by constrained sampling. International Journal of Intelligent Systems, 37(1):30–51.

Medveď, M., Jakubíček, M., and Kovář, V. (2016). English-french document alignment based on keywords and statistical translation. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 728–732.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546.

Minh-Cong, N.-H., Van-Vinh, N., and Le-Minh, N. (2023a). A fast method to filter noisy parallel data wmt2023 shared task on parallel data curation. In Proceedings of the Eighth Conference on Machine Translation, pages 359–365.

Minh-Cong, N.-H., Vinh, N. V., and Le-Minh, N. (2023b). A fast method to filter noisy parallel data WMT2023 shared task on parallel data curation. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, Proceedings of the Eighth Conference on Machine Translation, pages 359–365, Singapore. Association for Computational Linguistics.

Moon, H., Park, C., Koo, S., Lee, J., Lee, S., Seo, J., Eo, S., Jang, Y., Kim, H., Lee, H.-g., et al. (2023). Doubts on the reliability of parallel corpus filtering. Expert Systems with Applications, 233:120962.

Morin, E., Hazem, A., Boudin, F., and Loginova-Clouet, E. (2015). LINA: Identifying comparable documents from Wikipedia. In Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, pages 88–91, Beijing, China. Association for Computational Linguistics.

Munteanu, D. S. and Marcu, D. (2002). Processing comparable corpora with bilingual suffix trees. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pages 289–295.

Munteanu, D. S. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. Computational Linguistics, 31(4):477–504.

Nag, S., Kale, M., Lakshminarasimhan, V., and Singhavi, S. (2020). Incorporating bilingual dictionaries for low resource semi-supervised neural machine translation. arXiv preprint arXiv:2004.02071.

Nagao, H. and Tsujii, J. (1986). The transfer phase of the mu machine translation system. In Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics.

Nagao, M., Tsujii, J., Mitamura, K., Hirakawa, H., and Kume, M. (1980). A machine translation system from japanese into english-another perspective of mt systems. In COLING 1980 Volume 1: The 8th International Conference on Computational Linguistics.

Nagy, A., Lakatos, D. P., Barta, B., Nanys, P., and Ács, J. (2023). Data augmentation for machine translation via dependency subtree swapping. arXiv preprint arXiv:2307.07025.

Naranpanawa, R., Perera, R., Fonseka, T., and Thayasivam, U. (2020). Analyzing subword techniques to improve english to sinhala neural machine translation. International Journal of Asian Language Processing, 30(04):2050017.

Nastase, V. and 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), pages 22–39.

Nastase, V. and 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), pages 203–214.

Ni, J., Abrego, G. H., Constant, N., Ma, J., Hall, K., Cer, D., and Yang, Y. (2022). Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In Findings of the Association for Computational Linguistics: ACL 2022, pages 1864–1874.

Nissanka, L., Pushpananda, B., and Weerasinghe, A. (2020). Exploring neural machine translation for sinhala-tamil languages pair. In 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), pages 202–207. IEEE.

Novák, A., Tihanyi, L., and Prószéky, G. (2008). The metamorpho translation system. In Proceedings of the Third Workshop on Statistical Machine Translation, pages 111–114.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. Computational linguistics, 29(1):19–51.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In Proceedings of NAACL-HLT 2019: Demonstrations, pages 48–53.

Papavassiliou, V., Prokopidis, P., and Piperidis, S. (2016). The ilsp/arc submission to the wmt 2016 bilingual document alignment shared task. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 733–739.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318.

Peng, W., Huang, C., Li, T., Chen, Y., and Liu, Q. (2020). Dictionary-based data augmentation for cross-domain neural machine translation. arXiv preprint arXiv:2004.02577.

Popović, M. (2015). chrf: character n-gram f-score for automatic mt evaluation. In Proceedings of the tenth workshop on statistical machine translation, pages 392–395.

Popović, M. (2017). chrf++: words helping character n-grams. In Proceedings of the second conference on machine translation, pages 612–618.

Post, M. (2018a). A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Post, M. (2018b). A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Pramodya, A. (2023). Exploring low-resource neural machine translation for sinhala-tamil language pair. In Proceedings of the 8th Student Research Workshop associated with the International Conference Recent Advances in Natural Language Processing, pages 87–97.

Pramodya, A., Pushpananda, R., and Weerasinghe, R. (2020). A comparison of transformer, recurrent neural networks and smt in tamil to sinhala mt. In 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), pages 155–160. IEEE.

Priyadarshani, H., Rajapaksha, M., Ranasinghe, M., Sarveswaran, K., and Dias, G. (2019). Statistical machine learning for transliteration: Transliterating names between sinhala, tamil and english. In 2019 International Conference on Asian Language Processing (IALP), pages 244–249. IEEE.

Prószéky, G. (2005). An approach to machine translation via the rule-to-rule hypothesis. In Proceedings of the 10th EAMT Conference: Practical applications of machine translation.

Pushpananda, R. (2019). Improving sinhala-tamil translation through deep learning techniques.

Rajitha, M., Piyarathna, L., Nayanajith, M., and Surangika, S. (2020). Sinhala and english document alignment using statistical machine translation. In 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), pages 29–34. IEEE.

Rajpurkar, P., Zhang, J., Lopyrev, K., and 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, pages 2383–2392.

Ramesh, A., Parthasarathy, V. B., Haque, R., and Way, A. (2021a). Comparing statistical and neural machine translation performance on hindi-to-tamil and english-to-tamil. Digital, 1(2):86–102.

Ramesh, A., Uhana, H. U., Parthasarathy, V. B., Haque, R., and Way, A. (2021b). Augmenting training data for low-resource neural machine translation via bilingual word embeddings and bert language modelling. In 2021 International Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE.

Ranathunga, S. and 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, pages 823–848.

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

Ranathunga, S., Lee, E.-S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., and Kaur, R. (2023). Neural machine translation for low-resource languages: A survey. ACM Computing Surveys, 55(11):1–37.

Ranathunga, S., Lee, E.-S. A., Skenduli, M. P., Shekhar, R., Alam, M., and Kaur, R. (2021). Neural machine translation for low-resource languages: A survey. arXiv preprint arXiv:2106.15115.

Ranathunga, S., Ranasinghea, A., Shamala, J., Dandeniyaa, A., Galappaththia, R., and Samaraweeraa, M. (2024b). A multi-way parallel named entity annotated corpus for english, tamil and sinhala. arXiv preprint arXiv:2412.02056.

Rathnayake, H., Sumanapala, J., Rukshani, R., and Ranathunga, S. (2022). Adapter-based fine-tuning of pre-trained multilingual language models for code-mixed and code-switched text classification. Knowledge and Information Systems, 64(7):1937–1966.

Reimers, N., Gurevych, I., Reimers, N., Gurevych, I., Thakur, N., Reimers, N., Daxenberger, J., and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Resnik, P. (1998). Parallel strands: A preliminary investigation into mining the web for bilingual text. In Conference of the Association for Machine Translation in the Americas, pages 72–82. Springer.

Resnik, P. (1999). Mining the web for bilingual text. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pages 527–534.

Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. Computational Linguistics, 29(3):349–380.

Rossenbach, N., Rosendahl, J., Kim, Y., Graça, M., Gokrani, A., and Ney, H. (2018). The rwth aachen university filtering system for the wmt 2018 parallel corpus filtering task. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pages 946–954.

Roy, A., Ray, P., Maheshwari, A., Sarkar, S., and Goyal, P. (2024). Enhancing low-resource nmt with a multilingual encoder and knowledge distillation: A case study. In Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024), pages 64–73.

Sachintha, D., Piyarathna, L., Rajitha, C., and Ranathunga, S. (2021). Exploiting parallel corpora to improve multilingual embedding based document and sentence alignment. arXiv preprint arXiv:2106.06766.

San, M. E., Usanavasin, S., Thu, Y. K., and Okumura, M. (2024). A study for enhancing low-resource thai-myanmar-english neural machine translation. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(4):1–24.

Sánchez-Martínez, F., Perez-Ortiz, J. A., Galiano Jimenez, A., and Oliver, A. (2024). Findings of the WMT 2024 shared task translation into low-resource languages of Spain: Blending rule-based and neural systems. In Haddow, B., Kocmi, T., Koehn, P., and Monz, C., editors, Proceedings of the Ninth Conference on Machine Translation, pages 684–698, Miami, Florida, USA. Association for Computational Linguistics.

Sarikaya, R., Maskey, S., Zhang, R., Jan, E.-E., Wang, D., Ramabhadran, B., and Roukos, S. (2009). Iterative sentence-pair extraction from quasi-parallel corpora for machine translation. In Tenth Annual Conference of the International Speech Communication Association, pages 432–435.

Sarveswaran, K. and Dias, G. (2020). Thamizhiudp: A dependency parser for tamil. In Proceedings of the 17th International Conference on Natural Language Processing (ICON), pages 200–207.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2021a). Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1351–1361.

Schwenk, H., Wenzek, G., Edunov, S., Grave, É., Joulin, A., and Fan, A. (2021b). Ccmatrix: Mining billions of high-quality parallel sentences on the web. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6490–6500.

Sen, S., Hasanuzzaman, M., Ekbal, A., Bhattacharyya, P., and Way, A. (2021). Neural machine translation of low-resource languages using smt phrase pair injection. Natural Language Engineering, 27:271–292.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 86–96.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725.

Shi, L., Niu, C., Zhou, M., and Gao, J. (2006). A DOM tree alignment model for mining parallel data from the web. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pages 489–496.

Shi, S., Wu, X., Su, R., and Huang, H. (2022). Low-resource neural machine translation: Methods and trends. ACM Transactions on Asian and Low-Resource Language Information Processing, 21(5):1–22.

Shliazhko, A., Oguejiofor, A., Agafonova, A., et al. (2022). mgpt: Few-shot learners go multilingual. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 1492–1509.

Sloto, S., Thompson, B., Khayrallah, H., Domhan, T., Gowda, T., and Koehn, P. (2023). Findings of the wmt 2023 shared task on parallel data curation. In Proceedings of the Eighth Conference on Machine Translation, pages 95–102.

Stefanescu, D., Ion, R., and Hunsicker, S. (2012). Hybrid parallel sentence mining from comparable corpora. In Proceedings of the 16th Annual conference of the European Association for Machine Translation, pages 137–144.

Steingrímsson, S. (2023). A sentence alignment approach to document alignment and multi-faceted filtering for curating parallel sentence pairs from web-crawled data. In Proceedings of the Eighth Conference on Machine Translation, pages 366–374.

Steingrímsson, S., Loftsson, H., and Way, A. (2023). Filtering matters: Experiments in filtering training sets for machine translation. In Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), pages 588–600.

Stocke, A. (2011). Srilm at sixteen: Update and outlook. In Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Waikoloa, Hawaii, Dec. 2011.

Stojanovski, D. (2021). Modeling contextual information in neural machine translation. PhD thesis, lmu.

Su, T., Peng, X., Thillainathan, S., Guzmán, D., Ranathunga, S., and Lee, E.-S. (2024). Unlocking parameter-efficient fine-tuning for low-resource language translation. In Findings of the Association for Computational Linguistics: NAACL 2024, pages 4217–4225.

Sun, S., Zhuang, S., Wang, S., and Zuccon, G. (2025). An investigation of prompt variations for zero-shot llm-based rankers. In European Conference on Information Retrieval, pages 185–201. Springer.

Sun, Y., He, J., Xia, M., and Neubig, G. (2021). Contrastive learning for unsupervised neural machine translation. In ACL.

Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Tian, X., Zhu, D., Tian, H., and Wu, H. (2019). Ernie: Enhanced representation through knowledge integration. arXiv preprint arXiv:1904.09223.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems, 27:3104–3112.

Takase, S. and Kiyono, S. (2023). Lessons on parameter sharing across layers in transformers. In Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP), pages 78–90.

Tan, X., Ren, Y., He, D., Qin, T., Zhao, Z., and Liu, T.-Y. (2019). Multilingual neural machine translation with knowledge distillation. arXiv e-prints, pages arXiv–1902.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2021). Multilingual translation from denoising pre-training. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3450–3466.

Tars, M., Tattar, A., and Fishel, M. (2022). Cross-lingual transfer from large multilingual translation models to unseen under-resourced languages. Baltic Journal of Modern Computing, 10(3):435–446.

Tennage, P., Herath, A., Thilakarathne, M., Sandaruwan, P., and Ranathunga, S. (2018a). Transliteration and byte pair encoding to improve tamil to sinhala neural machine translation. In 2018 Moratuwa Engineering Research Conference (MERCon), pages 390–395. IEEE.

Tennage, P., Sandaruwan, P., Thilakarathne, M., Herath, A., and Ranathunga, S. (2018b). Handling rare word problem using synthetic training data for sinhala and tamil neural machine translation. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Tennage, P., Sandaruwan, P., Thilakarathne, M., Herath, A., Ranathunga, S., Jayasena, S., and Dias, G. (2017). Neural machine translation for sinhala and tamil languages. In 2017 International Conference on Asian Language Processing (IALP), pages 189–192. IEEE.

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

Thompson, B. and Koehn, P. (2019). Vecalign: Improved sentence alignment in linear time and space. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1342–1348.

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

Udawatta, P., Udayangana, I., Gamage, C., Shekhar, R., and Ranathunga, S. (2024). Use of prompt-based learning for code-mixed and code-switched text classification. World Wide Web, 27(5):63.

Uszkoreit, J., Ponte, J., Popat, A., and Dubiner, M. (2010). Large scale parallel document mining for machine translation. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1101–1109.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. (2007). Parallel corpora for medium density languages. Amsterdam Studies In The Theory And History Of Linguistic Science Series 4, 292:247.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30:5998–6008.

Velayuthan, M., Jayakody, D., De Silva, N., Fernando, A., and Ranathunga, S. (2024). Back to the stats: Rescuing low resource neural machine translation with statistical methods. In Proceedings of the Ninth Conference on Machine Translation, pages 901–907.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018a). 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, pages 353–355.

Wang, J., Lu, Y., Weber, M., Ryabinin, M., Adelani, D., Chen, Y., Tang, R., and Stenetorp, P. (2025). Multilingual language model pretraining using machine-translated data. arXiv preprint arXiv:2502.13252.

Wang, X., Pham, H., Dai, Z., and Neubig, G. (2018b). Switchout: an efficient data augmentation algorithm for neural machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 856–861.

Wang, Z., Wang, P., Liu, K., Wang, P., Fu, Y., Lu, C.-T., Aggarwal, C. C., Pei, J., and Zhou, Y. (2024). A comprehensive survey on data augmentation. arXiv preprint arXiv:2405.09591.

Weaver, W. (1955). Translation. In Machine Trans Languages, volume 14, pages 15–23.

Weller-Di Marco, M. and Fraser, A. (2022). Findings of the WMT 2022 shared tasks in unsupervised MT and very low resource supervised MT. In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, Proceedings of the Seventh Conference on Machine Translation (WMT), pages 801–805, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. Proceedings of the IEEE, 78(10):1550–1560.

Wettig, A., Gao, T., Zhong, Z., and 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, pages 2977–2992.

Winiwarter, W. (2007). Jetcat–japanese-english translation using corpus-based acquisition of transfer rules. JOURNAL OF COMPUTERS, 2(9):27.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021a). mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021b). mt5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), page 483–498.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G. H., Yuan, S., Tar, C., Sung, Y.-H., 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, pages 87–94.

Yang, Z., Li, Y., Liu, L., Li, R., and Li, M. (2023). Grammar-aware representation learning for unsupervised machine translation. In EMNLP.

Yazar, B. K., Şahın, D. Ö., and Kiliç, E. (2023). Low-resource neural machine translation: A systematic literature review. IEEE Access, 11:131775–131813.

Zafarian, A., Sadeghi, A. P. A., Azadi, F., Ghiasifard, S., Panahloo, Z. A., Bakhshaei, S., and Ziabary, S. M. M. (2015). Aut document alignment framework for bucc workshop shared task. In Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, pages 79–87.

Zhang, B., Nagesh, A., and Knight, K. (2020). Parallel corpus filtering via pre-trained language models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8545–8554.

Zhang, J., Luan, H., Sun, M., Zhai, F., Xu, J., Zhang, M., and Liu, Y. (2018). Improving the transformer translation model with document-level context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 533–542.

Zhang, J. and Zong, C. (2020). Neural machine translation: Challenges, progress and future. Science China Technological Sciences, 63(10):2028–2050.

Zhou, Y., Guo, C., Wang, X., Chang, Y., and Wu, Y. (2024). A survey on data augmentation in large model era. arXiv e-prints, pages arXiv–2401.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1. 0. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 3530–3534.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1568–1575.

Zweigenbaum, P., Sharoff, S., and Rapp, R. (2018). Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In Proceedings of 11th Workshop on Building and Using Comparable Corpora, pages 39–42.

# APPENDIX A

# MONOLINGUAL DATA FOR MLM STEP

The Table A.1 shows the results corresponding to Figure 5.4.

Table A.1: Bitext mining recall scores for using pure monolingual data versus source and target sides from a parallel corpus (as monolingual data) for MLM experiments.

| Dataset | Dataset Size | Army | | | Hiru | | | ITN | | | Newsfirst | | | Averages | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F | B | I | F | B | I | F | B | I | F | B | I | F | B | I |
| **Sinhala → English** | | | | | | | | | | | | | | | | |
| 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 |
| **Tamil → English** | | | | | | | | | | | | | | | | |
| 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

# PDC: EXTRINSIC EVALUATION RESULTS

The NMT evaluation scores for the parallel data curation task are reported in Table B.1 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 B.1: NMT scores on the Flores+ devtest using top 50,000 parallel sentences from the ranked NLLB and CCAligned corpus.

| | **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

# DEBIASING DISPARITY WITH HEURISTICS

## C.1 Improved Taxonomy for Noise Categorization

Table C.1 shows example parallel sentences which falls into the ***CCN*** noise category, which is newly added to the taxonomy. We show the final error taxonomy including definitions for each noise category in Table C.2.

Table C.1: Example parallel sentences which will be separately identified under the new noise category *CCN*

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

Table C.2: Ranathunga et al. (2024a)'s error taxonomy with the CCN category that has been newly added by us

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

## C.2 Human Evaluation

We conduct a human evaluation to quantify the noise in the top-ranked parallel corpora by the multiPLMs, before and after applying the heuristics. Table C.3 shows the years

of experience and the qualifications of those annotators who conducted this task.

Table C.3: Annotator details with the years of experience and their qualifications.

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

During training the annotation guideline, given in terms of a flowchart is shown in Figure C.1

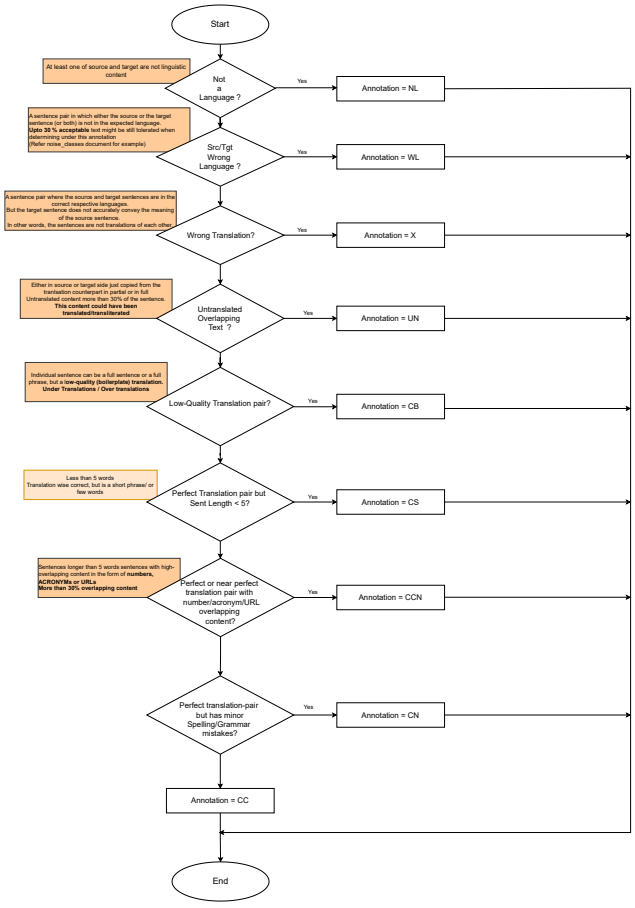


Figure C.1: Shows the annotation guideline document in terms of a flow chart. This shows the priority of the noise category to be selected prior to declaring the annotation class.

Finally, we show the final dataset sizes after applying the heuristic along with the percentage reduction in Table C.4.

Table C.4: Shows the final corpus sizes after applying heuristics, along with the reduction percentage. Here **DD+PN** is *Deduplication+punctNums*, **SL** is *sLength* and **LT** is *LIDThresh*. **NA** corresponds to the experiments that are not applicable for the language pair. **Red(%)** refers to the percentage reduction of the dataset size due to applying the heuristics.

| Heuristic | Applicable Side | Sinhala - Tamil | | | | English - Sinhala | | | | English - Tamil | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CCMatrix | | CCAligned | | CCMatrix | | CCAligned | | CCMatrix | | CCAligned | |
| | | Sents. | Red(%) | Sents. | Red(%) | Sents. | Red(%) | Sents. | Red(%) | Sents. | Red(%) | Sents. | Red(%) |
| Baseline | | 215965 | | 260119 | | 6270800 | | 619730 | | 7291118 | | 880568 | |
| DD | S | 189654 | | 250038 | 3.88% | 6146819 | 1.98% | 570768 | 7.90% | 6378607 | 12.52% | 797071 | 9.48% |
| | T | 209461 | | 247176 | 4.98% | 3242950 | 48.28% | 562088 | 9.30% | 4754106 | 34.80% | 780355 | 11.38% |
| | ST | 183904 | | 243384 | 6.43% | 3176145 | 49.35% | 537581 | 13.26% | 4060447 | 44.31% | 736212 | 16.39% |
| DD-4gram | S | 176590 | 18.23% | 189218 | 27.26% | 4171884 | 33.47% | 403993 | 34.81% | 4802654 | 34.13% | 440248 | 50.00% |
| | T | 196538 | 9.00% | 207603 | 20.19% | 2751819 | 56.12% | 423752 | 31.62% | 4271550 | 41.41% | 549621 | 37.58% |
| | ST | 172440 | 20.15% | 184159 | 29.20% | 2035282 | 67.54% | 355790 | 42.59% | | 100.00% | 365648 | 58.48% |
| DD-5gram | S | 188457 | 12.74% | 217733 | 16.29% | 4486108 | 28.46% | 481021 | 22.38% | 5104643 | 29.99% | 558504 | 36.57% |
| | T | 204045 | 5.52% | 226738 | 12.83% | 3071693 | 51.02% | 499307 | 19.43% | 4516819 | 38.05% | 638600 | 27.48% |
| | ST | 185915 | 13.91% | 216389 | 16.81% | 2374578 | 62.13% | 446838 | 27.90% | 3319964 | 54.47% | 487465 | 44.64% |
| DD-6gram | S | 196194 | 9.15% | 232604 | 10.58% | 5383674 | 14.15% | 528525 | 14.72% | 5309083 | 27.18% | 639961 | 27.32% |
| | T | 200310 | 7.25% | 237467 | 8.71% | 3142124 | 49.89% | 539513 | 12.94% | 4590987 | 37.03% | 681186 | 22.64% |
| | ST | 196194 | 9.15% | 233792 | 10.12% | 2859756 | 54.40% | 505429 | 18.44% | 3489629 | 52.14% | 570403 | 35.22% |
| DD-7gram | S | 200899 | 6.98% | 240704 | 7.46% | 5701801 | 9.07% | 554025 | 10.60% | 5718913 | 21.56% | 679750 | 22.81% |
| | T | 204485 | 5.32% | 244007 | 6.19% | 3170538 | 49.44% | 561285 | 9.43% | 4631486 | 36.48% | 703950 | 20.06% |
| | ST | 200898 | 6.98% | 246104 | 5.39% | 3021457 | 51.82% | 538898 | 13.04% | 3745771 | 48.63% | 611821 | 30.52% |
| DD+N | S | 182386 | 15.55% | 260119 | 0.00% | 6105433 | 2.64% | 505828 | 18.38% | 6337285 | 13.08% | 683882 | 22.34% |
| | T | 201551 | 6.67% | 218980 | 15.82% | 3225979 | 48.56% | 502971 | 18.84% | 4722644 | 35.23% | 675627 | 23.27% |
| | ST | 176040 | 18.49% | 216238 | 16.87% | 3158067 | 49.64% | 476379 | 23.13% | 4031459 | 44.71% | 631485 | 28.29% |
| DD+PN | S | 180380 | 16.48% | 215100 | 17.31% | 5931349 | 5.41% | 494778 | 20.16% | 6194331 | 15.04% | 668849 | 24.04% |
| | T | 198352 | 8.16% | 212341 | 18.37% | 3197186 | 49.01% | 492801 | 20.48% | 4666158 | 36.00% | 660902 | 24.95% |
| | ST | 173804 | 19.52% | 207197 | 20.35% | 3130297 | 50.08% | 465617 | 24.87% | 3987832 | 45.31% | 616702 | 29.97% |
| DD+PN+4gram | ST+T | | | | | | | 289248 | 53.33% | | | | |
| DD+PN+5gram | ST + T | 167022 | 22.66% | 187250 | 28.01% | 3044520 | 51.45% | 380146 | 38.66% | | | | |
| DD+PN+6gram | ST + T | 169784 | 21.38% | 189060 | 27.32% | | | 428939 | 30.79% | 4547759 | 37.63% | 464424 | 47.26% |
| DD+PN+7gram | T +T | | | 196221 | 24.56% | | | | | 4620008 | 36.64% | | |
| SL | S | 150094 | 30.50% | 188061 | 27.70% | 5088747 | 18.85% | 411474 | 33.60% | 6498956 | 10.86% | 595057 | 32.42% |
| | T | 100799 | 53.33% | 161363 | 37.97% | 3670963 | 41.46% | 377708 | 39.05% | 4267495 | 41.47% | 517516 | 41.23% |
| | ST | 96264 | 55.43% | 157978 | 39.27% | 3341564 | 46.71% | 348829 | 43.71% | 4134919 | 43.29% | 491207 | 44.22% |
| LID | S | 192377 | 10.92% | 241617 | 7.11% | 6200355 | 1.12% | 479589 | 22.61% | 7210848 | 1.10% | 669260 | 24.00% |
| | T | 186720 | 13.54% | 241863 | 7.02% | 6066681 | 3.26% | 575298 | 7.17% | 6800923 | 6.72% | 794143 | 9.81% |
| | ST | 178276 | 17.45% | 231619 | 10.96% | 6010065 | 4.16% | 457639 | 26.16% | 6743988 | 7.50% | 625281 | 28.99% |
| LT | S | 181470 | 15.97% | 227791 | 12.43% | 6120792 | 2.39% | 398272 | 35.73% | 6120793 | 16.05% | 564870 | 35.85% |
| | T | 172726 | 20.02% | 222290 | 14.54% | 5990169 | 4.48% | 546472 | 11.82% | 5990170 | 17.84% | 731484 | 16.93% |
| | ST | 162777 | 24.63% | 208644 | 19.79% | 5877142 | 6.28% | 377579 | 39.07% | 5877143 | 19.39% | 518010 | 41.17% |
| STRatio | - | 170168 | 21.21% | 229101 | 11.92% | 4293239 | 31.54% | 459473 | 25.86% | 4051888 | 44.43% | 679820 | 22.80% |
| sentWRatio | S | 199788 | 7.49% | 246908 | 5.08% | 6232528 | 0.61% | 546460 | 11.82% | 7231531 | 0.82% | 743624 | 15.55% |
| | T | 196812 | 8.87% | 250151 | 3.83% | 6198124 | 1.16% | 552798 | 10.80% | 7176111 | 1.58% | 745221 | 15.37% |
| | ST | 193989 | 10.18% | 245161 | 5.75% | 6177212 | 1.49% | 531963 | 14.16% | 7138854 | 2.09% | 717159 | 18.56% |
| sentCRatio | S | 212287 | 1.70% | 224031 | 13.87% | 6262297 | 0.14% | 594169 | 4.12% | 7252444 | 0.53% | 832611 | 5.45% |
| | T | 212877 | 1.43% | 218726 | 15.91% | 6261991 | 0.14% | 596346 | 3.77% | 7281050 | 0.14% | 851343 | 3.32% |
| | ST | 211661 | 1.99% | 215151 | 17.29% | 6257079 | 0.22% | 588310 | 5.07% | 7247298 | 0.60% | 826866 | 6.10% |
| **Combined Heuristics** | | | | | | | | | | | | | |
| DD+PN+ngram (SiTa-CCMatrix n=5, SiTa-CCAligned n= 7 EnSi-CCMatrix/CCAligned n=5, EnTa-CCMatrix n=7, EnTa-CCAligned n=6) | | | | | | | | | | | | | |
| +SL | T+ST | 117198 | 45.73% | 143919 | 44.67% | 2245307 | 64.19% | 240086 | 61.26% | 3462458 | 52.51% | 321003 | 63.55% |
| +LT | T+ST | 130831 | 39.42% | 162543 | 37.51% | 2880530 | 54.06% | 239144 | 61.41% | 2876818 | 60.54% | 306352 | 65.21% |
| +sentWRatio | T+S | 154028 | 28.68% | 170715 | 34.37% | 2993730 | 52.26% | 337634 | 45.52% | 3377988 | 53.67% | 427587 | 51.44% |
| +SL+LT | T+ ST | 99207 | 54.06% | 127284 | 51.07% | 2200195 | 64.91% | 180731 | 70.84% | 4330140 | 40.61% | 241679 | 72.55% |
| +SL+sentCRatio | T+ST+ST | 116344 | 46.13% | 127035 | 51.16% | 2241513 | 64.25% | 224542 | 63.77% | 2794637 | 61.67% | 298141 | 66.14% |
| +SL+LT+sentWRatio | T+ST+ST+S | 127188 | 41.11% | 117970 | 54.65% | 2197629 | 64.95% | 177711 | 71.32% | 2726087 | 62.61% | 237105 | 73.07% |
| +SL+LT+sentWRatio>0.8 | T+ST+ST+ST | 179984 | 16.66% | 99311 | 61.82% | 2149037 | 65.73% | 161869 | 73.88% | 2203591 | 69.78% | 214639 | 75.62% |
| +SL+LT+sentCRatio | T+ST+ST+ST | 98894 | 54.21% | NA | | NA | | NA | | NA | | NA | |
| +SL+LT+STRatio | T+ST+ST+STR | 82866 | 61.63% | NA | | NA | | NA | | 2129744 | 70.79% | NA | |