Improving Data Augmentation Techniques to Generate Quality Parallel Data for Neural Machine Translation

WASA Fernando (208035D)

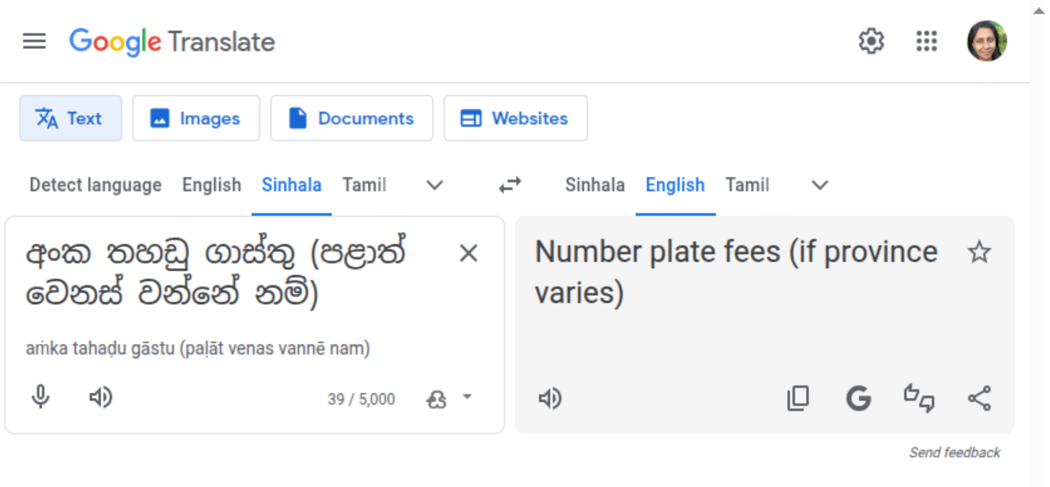
Supervisors

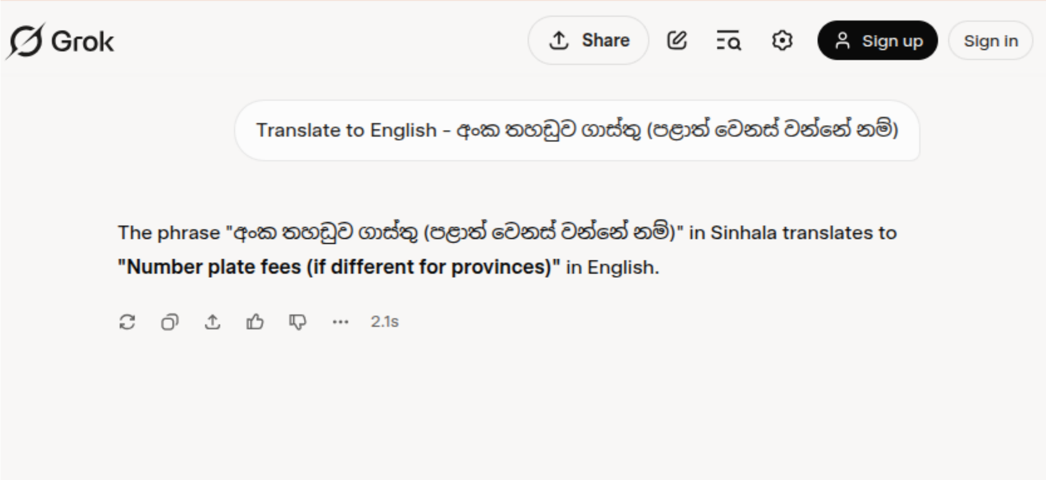
Dr Nisansa de Silva, Senior Lecturer, Depart. Computer Science & Engineering, University of Moratuwa
Dr Surangika Ranathunga, Senior Lecturer, Massey University, New Zealand

Motivation – Machine Translation

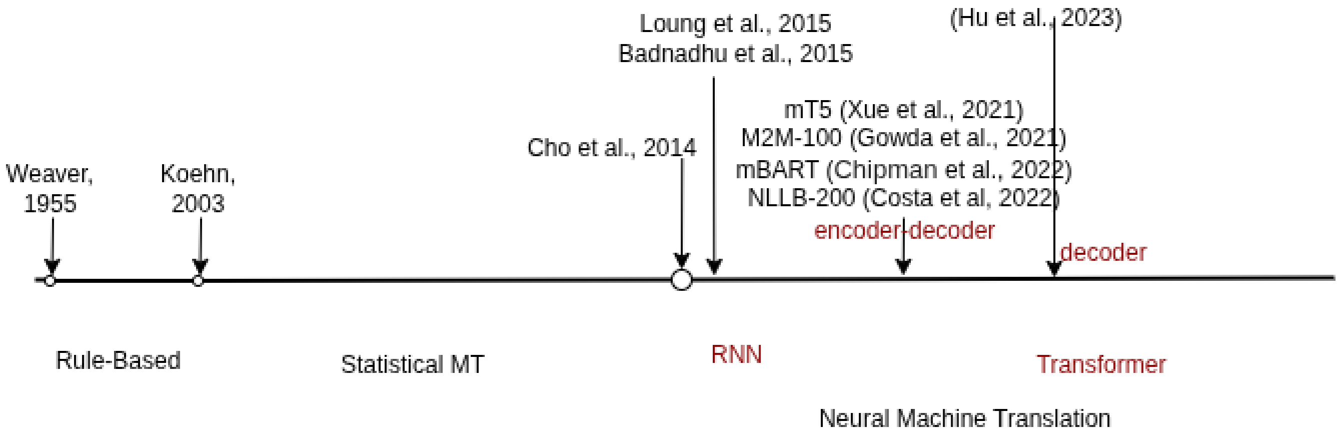
නීති කෙටුම්පත් සම්පාදක දෙපාර්තමේන්තුව
Legal Draftsman's Department

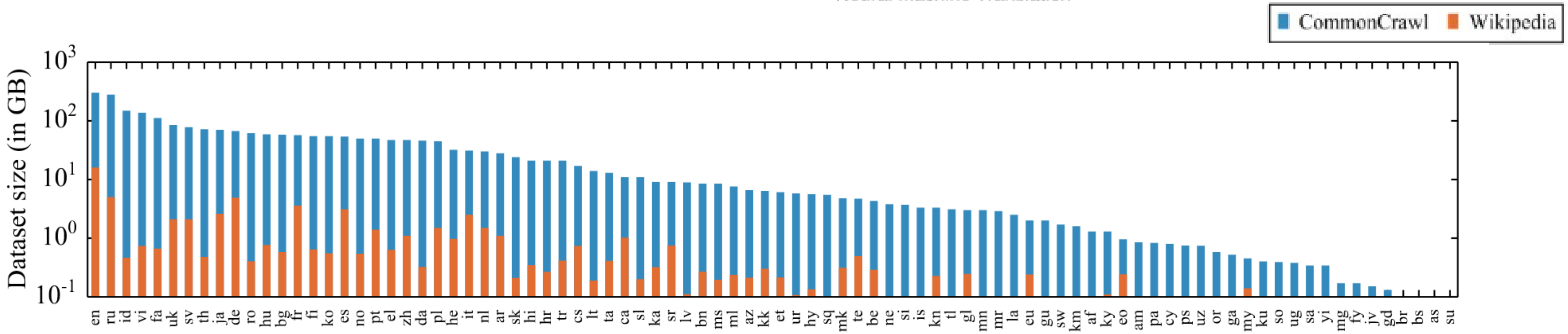
රාජ්‍ය භාෂා දෙපාර්තමේන්තුව
அரசகரும மொழிகள் திணைக்களம்
Department of Official Languages





Motivation – Timeline of Machine Translation





Language-wise training data distribution for XLM-R multiPLM (Conneau et al., 2020)

Weaver, W. (1952). Translation. In *Proceedings of the conference on mechanical translation*.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Luong, M. T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2020). mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.Gowda, T., Zhang, Z., Mattmann, C., & May, J. (2021, August). Many-to-English Machine Translation Tools, Data, and Pretrained Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations* (pp. 306-316).

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: BAYESIAN ADDITIVE REGRESSION TREES. *The Annals of Applied Statistics*, 266-298.

Costa-Jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., ... & NLLB Team. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Ruder, S., Søgaard, A., & Vulić, I. (2019, July). Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts* (pp. 31-38).

Xu, H., Kim, Y. J., Sharaf, A., & Awadalla, H. H. (2023). A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Terminology

- **Parallel Sentence-pair**

Source Sentence	Target Sentence
[en] Conducting Assistant Physiotherapist and Massage Certificate Course	[si] සහයක භෞත චිකිත්සක හා සම්බාහක සහතික පත්‍ර පාඨමාලාව පැවැත්වීම
[en] Ensuring compliance with the financial rules and regulations of the Government	[ta] அரசாங்கத்தின் நிதி விதிகள் மற்றும் ஒழுங்கு விதிகளுடன் இணங்கிச் செயற்படுதலை உறுதிப்படுத்தல்.

- **Supervised Neural Machine Translation (NMT)**
  - Given parallel sentences Neural Network learns to output a translation in the target language, given a sentence in the source language

- **High-Resource Languages (HRLs) vs Low Resource Languages (LRLs)**
  - Based on the dataset and linguistic resource availability (Joshi et al., 2020)

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020, July). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282-6293).

Motivation – Why NMT is challenging?

- Supervised NMT models trained on large parallel datasets using transformer architecture (Vaswani et al., 2017) produce state-of-the-art results (Haddow et al., 2022)
- When parallel data is limited, results for the same architectures **NMT results are suboptimal.**
  - Limited vocabulary coverage in the parallel data
  - Limited coverage of vocabulary in different contexts in the parallel data
  - In-adequate sequence-to-sequence mappings in the parallel corpus
- For morphologically rich languages, words inflect due to gender, number, case categories etc. leading to more vocabulary.
- Low-resource, morphologically rich languages the parallel data scarcity problem worsens the NMT performance

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.
Haddow, B., Bawden, R., Miceli-Barone, A. V., Helcl, J., & Birch, A. (2022). Survey of Low-Resource Machine Translation. *Computational Linguistics*, *48*(3), 673-732.

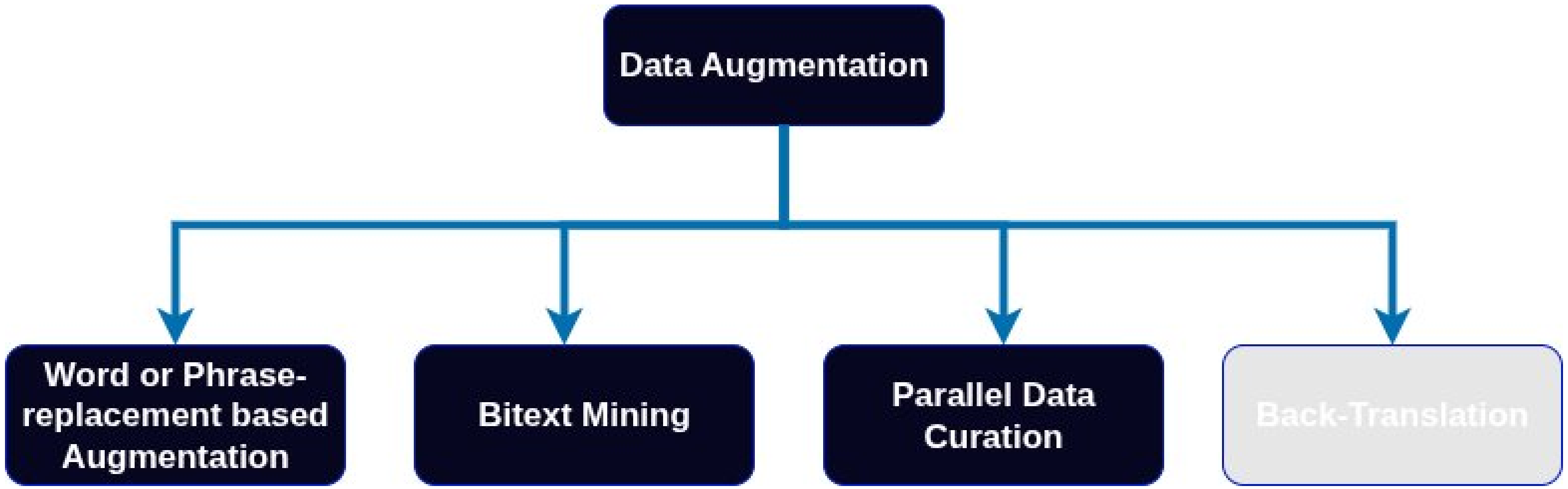
Motivation – Data Scarcity Problem for LRLs

- High-resource languages have large scale gold-standard parallel datasets.
  - **Europarl Parallel Corpus (Koehn, 2005)** - 1~2 Million sentences for high-resource languages.
  - **UN Parallel Corpus (Ziemski et al., 2016)** manual translations for 6 languages with minimum 16 Million sentences for each language.
- For Low resource languages s.a. Sinhala and Tamil, such gold-standard parallel datasets are in the range of 100k.
- **Parallel data scarcity problem** is a hindrance to the progress of NMT research among Sinhala-English-Tamil languages.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers* (pp. 79-86).
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.

Background

- **Data Augmentation** aims at alleviating the data scarcity problem by inducing parallel data synthetically or by automatic means.
- Data augmentation techniques categorization (Costa-jussà et al.,2022; Ranathunga et al., 2021)



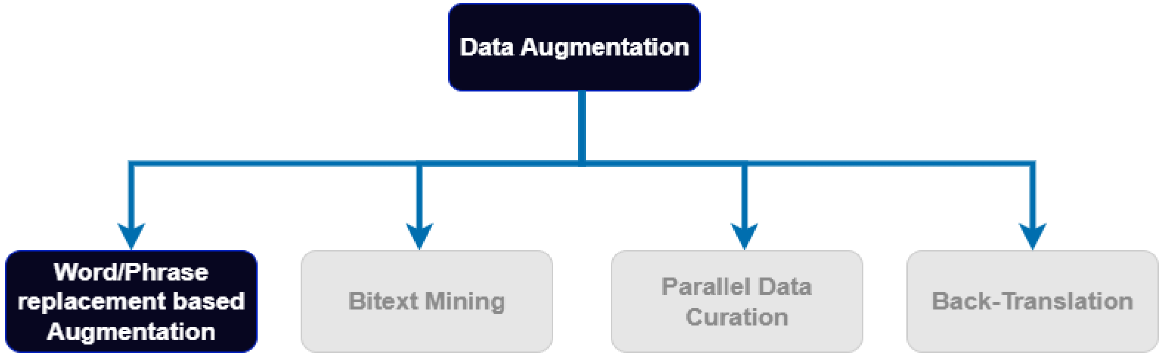
Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al. (2022). No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.
Ranathunga, S., Lee, E.-S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., and Kaur, R. (2023). Neural machine translation for low-resource languages: A survey. ACM Computing Surveys , 55(11):1–37.

Data Augmentation to Induce High-Quality Parallel Sentences for Low-Resource NMT

Gap	- Existing methods limited to a single OOV Type; either rare words or unseen words from a dictionary. - Existing methods limited to validating the synthetic sentence-pair either syntactically or semantically.	- No Empirical study to analyse the effectiveness of commonly used Multilingual Pre-trained Language Models (multiPLMs) for Document Alignment and Sentence Alignment Tasks for Low-resource setting.	- Encoder-based multiPLMs produced embeddings have weak cross-lingual alignment, especially for LRLs. Hence they perform poorly for sentence-retrieval tasks.	- The choice of multiPLMs in the Parallel Data Curation (PDC) task, leads to a disparity among NMT scores. - Lacks noise class in existing error taxonomy to identify noise introduced due to bias in multiPLMs.
Research Objectives	**RO1.** Implement an algorithm to generate synthetic parallel sentences to augment OOV terms.	**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	**RO3:** Improve the cross-lingual representations of existing multiPLMs to obtain High-Quality parallel sentences from the parallel sentence alignment task.	**RO4.** Exploring Parallel Data Curation (PDC) techniques to extract high-quality parallel sentences from web-mined parallel corpora
Contribution	- Algorithm to generate synthetic parallel sentences by augmenting OOV terms, by imposing both syntactic and semantic features to validate. - Publicly release the synthetic parallel sentences	- multiPLMs, trained using parallel data during the pre-training stage, are favourable for bitext-mining task for LRLs. - Publicly release the gold-standard human-annotated benchmark evaluation datasets for the Document and Sentence Alignment Tasks	- introduce an objective masking strategy termed Linguistic Entity Masking (LEM), to improve the cross-lingual representations of existing multiPLMs. - Empirical study on existing masking strategies - Publicly release cross-lingual improved multiPLM.	-Empirically find heuristic-combination leading to optimal NMT results and on the disparity among NMT models using multiPLM ranked parallel data. -Improve existing taxonomy and conduct a comparative human evaluation to quantify noise before and after heuristic-based filtration. -Publicly release curated datasets
Publication	Data Augmentation to Address Out of Vocabulary Problem in Low Resource Sinhala English Neural Machine Translation. ***PACLIC (2021)***	Exploiting bilingual lexicons to improve multilingual embedding-based document and sentence alignment for LRLs ***Knowledge and Information Systems (2023)***	LEM to Improve Cross-Lingual Representation of multiPLMs for Low-Resource Languages ***Knowledge and InformationSystems (2025)***	Improving the quality of Web-mined Parallel Corpora of Low-Resource Languages using Debiasing Heuristics. ***EMNLP (2025)***

RO1.
Implement an algorithm to generate synthetic parallel sentences to augment OOV terms.

Generating Synthetic Parallel Sentences



RO1: Motivation

- Generating Synthetic parallel sentences follows a word/phrase replacement approach
- Words to augment **Out-of-Vocabulary (OOV)**.
  - Rare Words (Tannage et al., 2018, Fadaee et al., 2017)
  - Unseen words, using a dictionary (Peng et al., 2022)
- Fadaee et al (2017) augment rare words and Tannage et al. (2018) improves this by validating with Part-of-Speech and morphological agreement.
- Peng et al (2020) augments out-of-domain dictionary and validates semantic agreement only.
- Substituting sub-trees (Alam et al., 2024) or top-most word (Duan et al., 2020) from dependency parser validates sentences syntactically.

Hypothesis : Use Syntactic and Semantic constraints to ensure syntactic and semantic correctness of synthetic parallel sentences.

Fadaee, M., Bisazza, A., & Monz, C. (2017, July). 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)* (pp. 567-57

Tennage, P., Sandaruwan, P., Thilakarathne, M., Herath, A., & Ranathunga, S. (2018, May). 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)*.

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

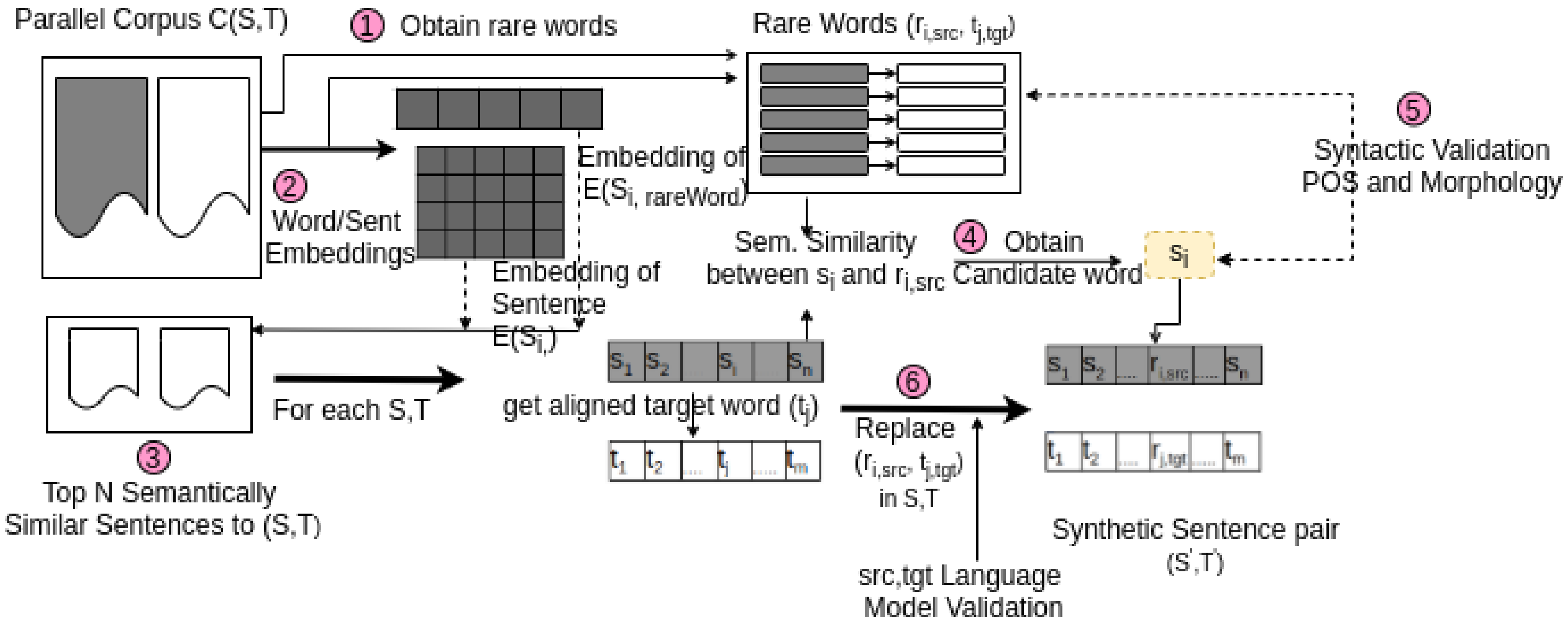
Alam, M. M. I., Ahmadi, S., and Anastasopoulos, A. (2024). A morphologically-aware dictionary-based data augmentation technique for machine translation of under-represented languages. arXiv preprint arXiv:2402.01939 .

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

RO1: Methodology – Rare word/Dictionary Augmentation

1

rare words (freq. = 1) from Source side.
Translation of rare word
aligned parallel sentence



2

Obtain word/sent embeddings

3

sentSim
Select Candidate pairs

4

wordSim
Select Candidate word for replacement

5

Syntactic constraints
POS
Morphology agreement

6

Tri-gram LM score
Synthetic source and
Synthetic target

RO1: Assumptions / Design Decisions

1. Augment both types of OOV – Rare words and Unseen words (dictionary terms)
2. Improving word embeddings to determine semantic similarity.
   - Post-processing of the word embeddings was done to improve the semantic similarity between the words.

     Eg: **run - running** vs **sing - chant**
   - Follow the work of Artexte et al (2018) and conducts a linear transformation on the word embeddings using an alpha (α) value.
3. Validations done to preserve syntactic and semantic correctness?
   - **Syntactic** constraints – POS and Morphology agreement
   - **Semantic** constraints – Sentence Similarity and Word Similarity
   - **Context** validation **–** Tri-gram replaced context is validated using Language Model

Artetxe Zurutuza, M., Labaka Intxauspe, G., López Gazpio, I., & Agirre Bengoa, E. (2018). Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation. In *The 22nd Conference on Computational Natural Language Learning: Proceedings of the Conference, October 31-November 1, 2018 Brussels, Belgium*. ACL.

RO1: Synthetic Parallel Sentence-pair

Rare Word /Translation	**පාර්ශ්වයන්**-----**parties**
Original source sent.	දිස්ත්‍රික් පරිපාලනයට හා ප්‍රාදේශීය පරිපාලනයට අදාළ **නිලධාරීන්** සම්බන්ධව ලැබෙන පෙත්සම් සහ පැමිණිලි සම්බන්ධව අපක්ෂපාතී පරීක්ෂණ පැවැත්වීම මඟින් යහපත් පාලනයක් ඇති කිරීම
Original target sent.	Creating better governance through conducting impartial investigation regarding petitions , complaints received in connection with relevant **officers** to District administration and Divisional administration
Synthetic source sent.	දිස්ත්‍රික් පරිපාලනයට හා ප්‍රාදේශීය පරිපාලනයට අදාළ **පාර්ශ්වයන්** සම්බන්ධව ලැබෙන පෙත්සම් සහ පැමිණිලි සම්බන්ධව අපක්ෂපාතී පරීක්ෂණ පැවැත්වීම මඟින් යහපත් පාලනයක් ඇති කිරීම
Synthetic target sent.	Creating better governance through conducting impartial investigation regarding petitions , complaints received in connection with relevant **parties** to District administration and Divisional administration

RO1: Experimental Setup

- Conducted Experiments for Sinhala-English language pair
- Datasets

Parallel Data	Traing Sentences	Validation Sentences
No. Sentences	54914	1623
No. of Words (En)	553002	23578
No. of Words (Si)	535185	22721

Government domain (Fernando et al., 2020)

Monolingual Data	English	Sinhala
No of Sentences	1.2 Million	1.2 Million
No of Words	51.1 Million	48.2 Million

Monolingual data (Isuranga et al., 2020)

	No of Sentences	No of Words / Unique Words		No of Rare Words		No of Dictionary Terms	
		Sinhala	English	Sinhala	English	Sinhala	English
Testset 01 (SITA-Eval)	1603	18513/4520	19248/4237	76	55	11	58
Testset 02 (Government)	1462	28918/5341	30437/4956	133	55	17	108
Testset 03 (Government)	1438	26308/5057	27815/4865	127	68	23	99

RO1: Experimental Setup

- **Dictionary** English-Sinhala in-house dictionary with 23660 terms
- Linguistic Tools / libraries used:

Word alignment	GIZA++ (Och and Ney, 2003)
PoS Tagger	English[1] and Sinhala TnT (Fernando et al., 2018)
Morphological Analyser	Sinmorphy (Kumarasinge et al., 2021)
word embeddings	Fasttext (Bojanowski et al., 2016)
Language Model	SRILM Toolkit[2]

- NMT Architecture - RNN encoder-decoder architecture with attention (Bahdanau et al., 2015)
- Evaluation metric BLEU (Papineni et al., 2001) scores.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. Computational linguistics , 29(1):19–51.
Kumarasinghe, K., Dias, G., & Herath, I. (2021, July). Sinmorphy: A morphological analyzer for the sinhala language. In *2021 Moratuwa Engineering Research Conference (MERCon)* (pp. 681-686). IEEE.
Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606.
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.

[1]https://spacy.io/
[2] http://www.speech.sri.com/projects/srilm/

RO1: Experiments & Results

Rare Word Augmentation

Experiment	Aug. Sent.	Si → En (BLEU)			Aug. Sent.	En → Si (BLEU)		
		TS1	TS2	TS3		TS1	TS2	TS3
Baseline [train54K]	-	<u>22.47</u>	<u>21.22</u>	<u>26.82</u>	-	<u>20.61</u>	<u>19.33</u>	<u>24.97</u>
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$_{wo\ pp}$	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**

- Rare Word Augmentation Gain(max)
  +0.91 Si→En /+0.74 En→Si
- Best scores when combining syntactic and semantic constraints. They exceed baseline scores
- Combining all constraints did not produce the best gains for Si -> En direction. Limitations with morphological analyser (similar pattern PoS, WordSim+PoS)
- SentSim+wordSim vs pos+morph produce **comparable** results

RO1: Experiments & Results

Dictionary Word Augmentation

Experiment	Aug. Sent.	Si → En(BLEU) Testset1	Testset2	Testset3	Aug. Sent.	En → Si (BLEU) Testset1	Testset2	Testset3
Baseline[train54K]		22.47	21.22	26.82		20.61	19.33	24.97
Baseline(Fadaee)	35901	21.59	19.36	22.70	49211	20.31	17.59	22.39
Baseline(Peng)	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**

- Dictionary Term Augmentation
  Gain(max) +0.18 Si→En / +0.71 En→Si
- In Si side dictionary terms as OOV in test sets were less (TS1-11 | TS2-17 | TS3-23). Therefore gains marginal.
- En->Si direction augmentation is effective.
- SentSim+wordSim vs pos+morph produce comparable results

RO1: Experiments & Results – Qualitative Analysis

Rare word	පරිශීලනය (parisílanaya)
Si Sentence	විනිශ්චයකාරවරුන්ගේ පරිශීලනය පිණිස පුස්තකාලය සඳහා 'නීතිය' පිළිබඳ නව ග්‍රන්ථ මිල දී ගන්නා ලදි *viniścayakāravarungē* ***pariśilanaya*** *piṇisa pustakālaya sañdahā 'nītiya' piḷibaňda 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.

Fluency and Accuracy of the Translation is improved with syntactic constraints and semantic constraints.

RO1: Limitations & Future Work

Limitations	Future Work
The Sinhala linguistic tools (PoS Tagger, Morphological Analyser, Alignment Tool) limitations	Re-evaluate the upon availability of better performing POS Taggers, morphological analysers
Used static word embeddings.	Instead of static embeddings using contextualized embeddings. Eg:sinBERT (Dananjaya et al., 2022)
Context validation using tri-gram statistical LM.	Determine sentence fluency using a Neural based language model.

Dhananjaya, V., Demotte, P., Ranathunga, S., & Jayasena, S. (2022, June). BERTifying Sinhala-A Comprehensive Analysis of Pre-trained Language Models for Sinhala Text Classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 7377-7385).
Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022, May). Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 878-891).

RO1: Contributions & Publication

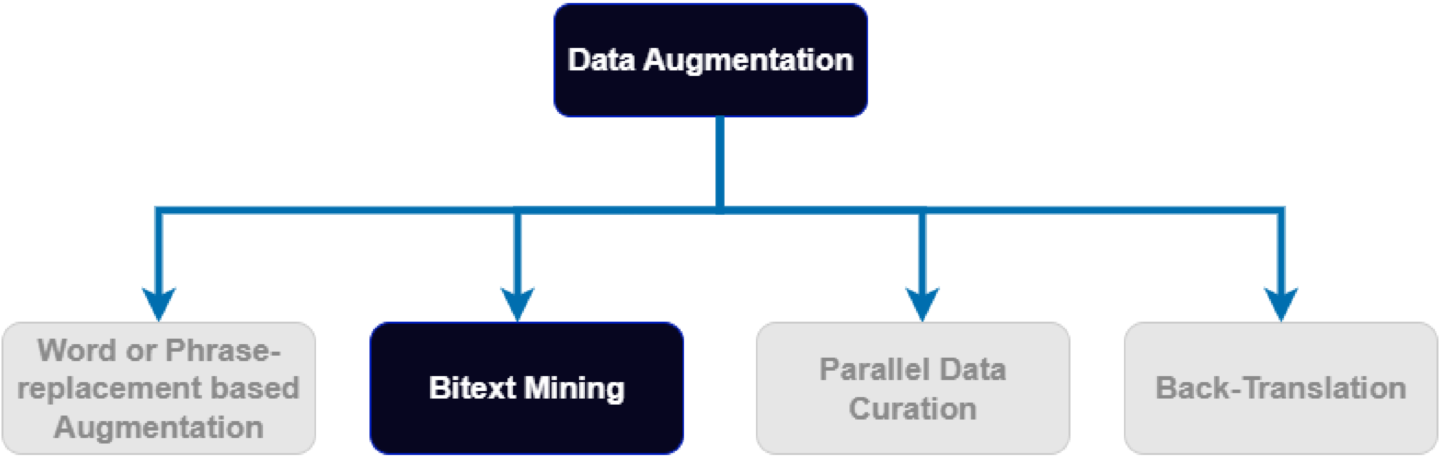
- Introduce an objective masking strategy termed Linguistic Entity Masking (LEM), to improve the cross-lingual representations of existing multiPLMs.
- This has been done using sentences from a parallel corpus with 56K only.
  Hence favourable for LRLs
- Publicly release the improved encoders for En-Si, En-Ta and Si-Ta language-pairs.

Publication

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). **(PACLIC,2021)**
h5-Index: 13

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

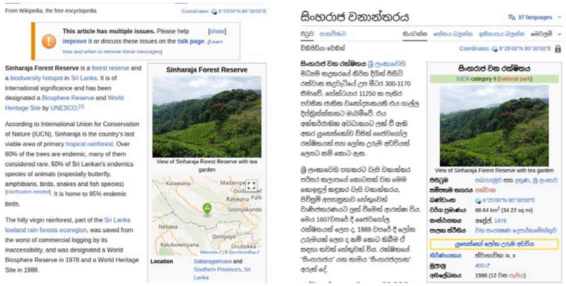
Empirical Study : multiPLMs for Bitext mining



RO2: Empirical Study using multiPLMs for Bitext Mining - Motivation

- The web contains human-created text in multiple languages at scale even for low-resource languages.
- Considering content availability in multiple languages, parallel sentences can be identified – **Bitext mining**.
- Shared tasks have taken place to encourage research in this direction BUCC2015-2018, 2024[1] and WMT2016-2020[2].
- Bitext mining pipeline
    - Identify & Crawl Web Data
    - Document Alignment
    - Sentence Alignment
    - Parallel Sentence Filtration

[1]https://comparable.limsi.fr/
[2]https://www2.statmt.org/wmt25/

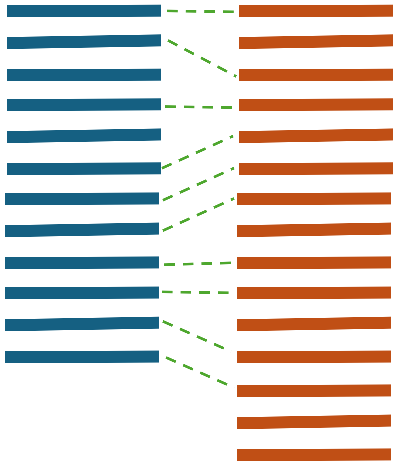


https://en.wikipedia.org/wiki/Sinharaja_Forest_Reserve
https://si.wikipedia.org/wiki/සිංහරාජ_වනාන්තරය

Document Alignment

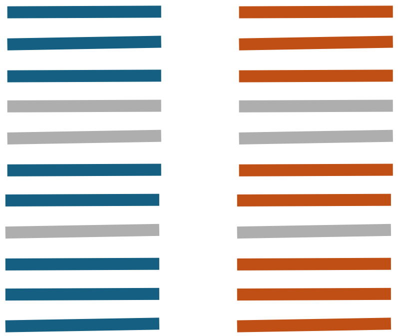
Identify & Crawl Web Pages

English Sinhala



Sentence Alignment

English Sinhala



Parallel Data Curation

English Sinhala

Parallel Sentences

RO2: Bitext Mining – Related Work

Document Alignment and Sentences alignment tasks are critical to determine the quality of the parallel sentences

Feature-based	URL (Resnik et al., 1999)
	DOM Tree alignment model. Textual content by means of HTML document structure. (Shi et al., 2006)
Machine Translation-based	Translating target to source and vice versa and measure similarity (Uszkoreit et al., 2010)
Vectorizerizing	Vectorizing considering bi-gram (Dara and Lin., 2016) and determine similarity by means of cosine sim.
Embedding Based	Similarity between document embeddings derived from sentence embeddings (Guo et al., 2019)
	Uses LASER2 to determine document similarity (El-kishky and Guzman, 2020)

RO2: Bitext Mining – Related Work

Document Alignment and Sentences alignment tasks are critical to determine the quality of the parallel sentences

Feature-based	Scoring functions with characters or words (Brown et al.,1991; Gale and Church., 1993)
Machine Translation-based	Uses phrase tables from statistical MT system(Gomes and Lopes, 2016)
Embedding Based	Vecalign uses bilingual embeddings (Thompson and Koehn, 2019)
	Pre-trained LASER2 embeddings (Bañón et al., 2020)
	Uses unsupervised multilingual embeddings for sentence alignment (Kvapilíková et al., 2020)
	Margine-based cosine similarity over LASER2 embeddings (Artetxe and Schwenk, 2019)

RO2: Research Questions

What characteristics in multiPLMs are influential for document alignment and sentence alignment tasks?

multiPLM	Architecture	Training Data	Pre-training/fine-tuning
LASER2 (Artetxe and Schwenk, 2019)	LSTM	parallel	Pre-training
XLM-R (Conneau et al., 2020)	Transformer	mono	Pre-Training
LaBSE (Feng et al., 2022)	Transformer	Mono + parallel	Pre-Training + Fine-tuning

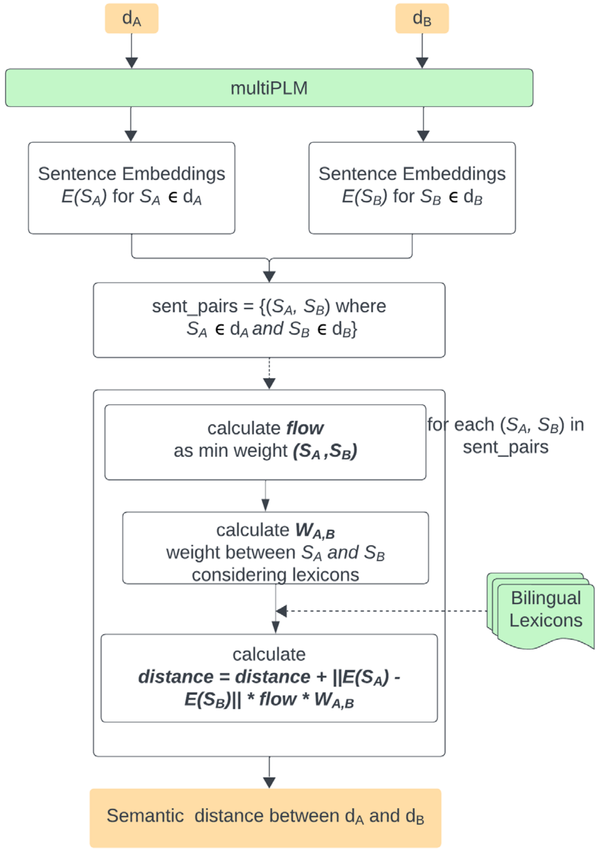
Can improvements using bilingual lexicons improve these results further?

Uses bilingual lexicons to improve the semantic similarity score in determining the document similarity and the sentence similarity (Rajitha et al., 2020)

RO2: Methodology

- Extended gold-standard evaluation benchmark dataset by Rajitha et al. (2020) for document alignment and sentence alignment tasks.
- Conducted intrinsic evaluation for document alignment and sentence alignment using the compiled gold-standard evaluation set.
- Evaluated the significance of bilingual lexicon based improvement (by means of a weighting) to the distance calculation function by Rajitha et al.(2020)
- Conducted extrinsic evaluation by training NMT systems for the six directions (Si→En, En→Si, Ta→En, En→Ta, Si→Ta, Ta→Si)

RO2: Document Alignment Algorithm (El-kishky and Guzman, 2020)



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

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

Distance calculation between two sentences

$$distance = distance + \|s_A - s_B\| \times flow$$

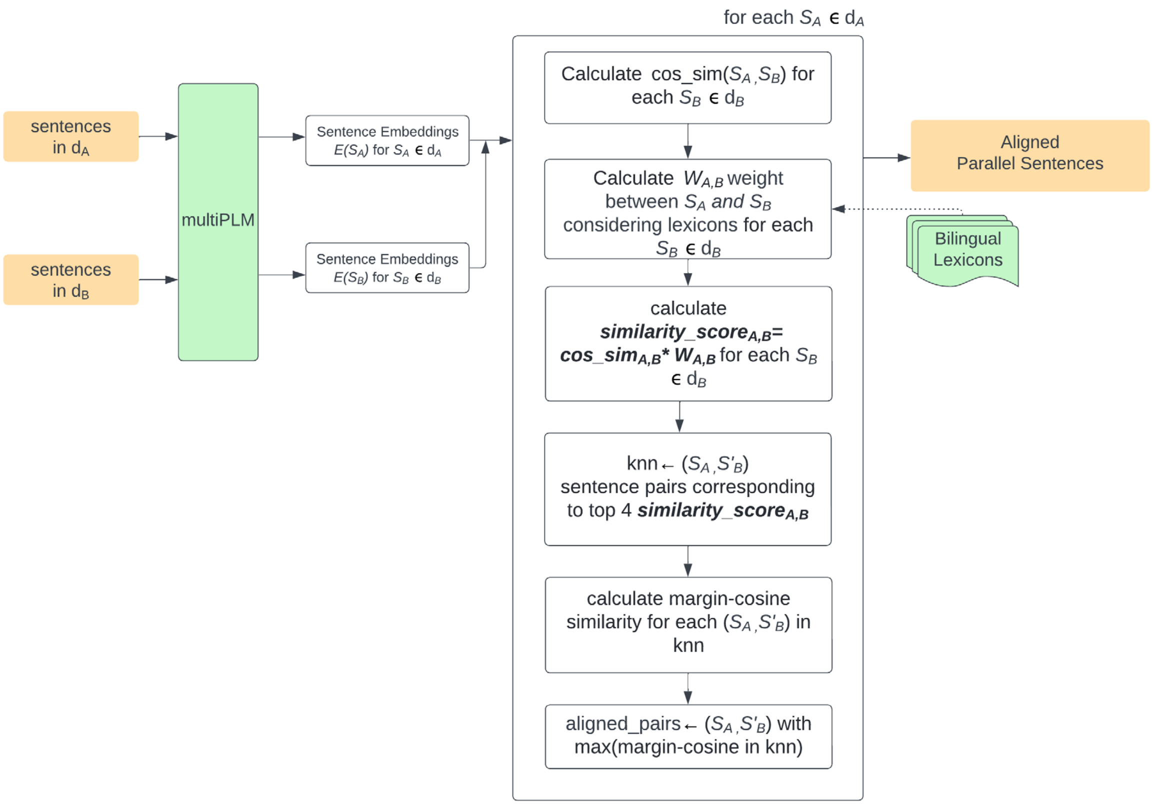
New weighting Scheme (Rajitha et al., 2020)

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

Modification to the distance calculation

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

RO2: Sentence Alignment Algorithm (Artetxe and Schwenk, 2019)



Improvement for the distance calculation (Rajitha et al., 2020)

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

Weighting Scheme (Rajitha et al., 2020)

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

RO2 : Experiments & Results : Document Alignment

En-Si

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

En-Ta

Experiment	Wt.	Hiru			ITN			Newsfirst			Army		
		R	P	F1	R	P	F1	R	P	F1	R	P	F1
LASER													
BL	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
BL+N	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
BL+N+Ds	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
BL+N+Ds+Dc	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
BL+N+Ds+MDc	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													
BL	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
BL+N	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
BL+N+Ds	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
BL+N+Ds+Dc	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
BL+N+Ds+MDc	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													
BL	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
BL+N	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

Si-Ta

Experiment	Wt.	Hiru			ITN			Newsfirst			Army		
		R	P	F1	R	P	F1	R	P	F1	R	P	F1
LASER													
BL	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
BL+N	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
BL+N+Ds	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
BL+N+Ds+Dc	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
BL+N+Ds+MDc	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													
BL	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
BL+N	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
BL+N+Ds	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
BL+N+Ds+Dc	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
BL+N+Ds+MDc	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													
BL	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**
BL+N	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**

- LASER2 baseline outperformed significantly with dictionary improvement **En-Ta +44%, Si-Ta +13% and Si-En +2%**
- LaBSE and XLM-R outperform LASER2 results for all three language-pairs.

RO2 : Experiments & Results : Sentence Alignment

	Experiments	Army						Hiru						ITN						Newsfirst					
		Forward		Backward		Intersection		Forward		Backward		Intersection		Forward		Backward		Intersection		Forward		Backward		Intersection	
		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 [14]		4352					29.00	1650					11.63	688					9.00	576					11.00
LaBSE [18]	BL	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**
	BL+Dict	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**
Laser	BL	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
	BL+Dict	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	BL	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
	BL+Dict	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	BL	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**
	BL+Dict	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 [18]	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
	BL+Dict	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
Laser	BL	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
	BL+Dict	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	BL	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
	BL+Dict	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	BL	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
	BL+Dict	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 [18]	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
	BL+Dict	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**
Laser	BL	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
	BL+Dict	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	BL	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
	BL+Dict	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	BL	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
	BL+Dict	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**

- Embeddings obtained from LaBSE performing best then XLM-R and LASER2
- Here the Dictionary improvement was less significant with LaBSE.

RO2 : Experiments & Results : NMT Experiments

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

- NMT scores are low due to the lack of training dataset size. Ie. EnSi, 25k~17k, EnTa 17k~13k and SiTa 21k~13k
- Improving distant scoring function (bilingual lexicons) has an impact to improve NMT results.
- LASER2 and LaBSE performed well in NMT compared to XLM-R.
- Using parallel data in pre-trainingor in fine-tuning stages in the multiPLM is favourable to produce quality parallelsentences,compared to multiPLM undergoing purely monolingual data.

RO2: Limitations & Future Work

Limitations	Future Work
We consider only encoder-based multiPLMs	Extend this study using encoder-decoder based sequence-to-sequence models (Ni et al., 2022) and decoder-based generative LLMs (Sun et al., 2025) used for cross-lingual sentence retrieval tasks.

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.
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.

RO2: Contributions & Publication

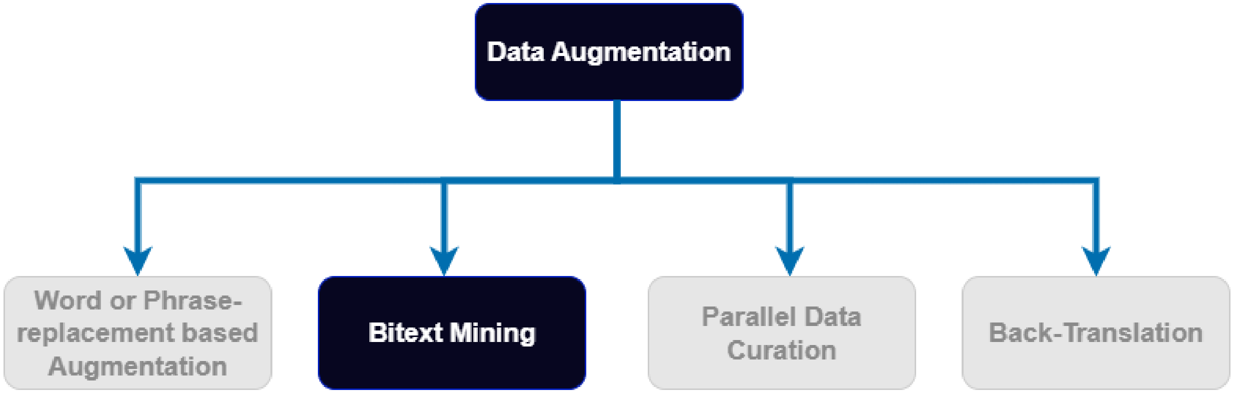
- From empirical study, identifying that pre-trained models which had undergone continual pre-training with parallel data perform well for document alignment and sentence alignment tasks.
- Release the extended document alignment and sentence alignment evaluation set, which was initially done by Rajitha et al., (2020)

Publication

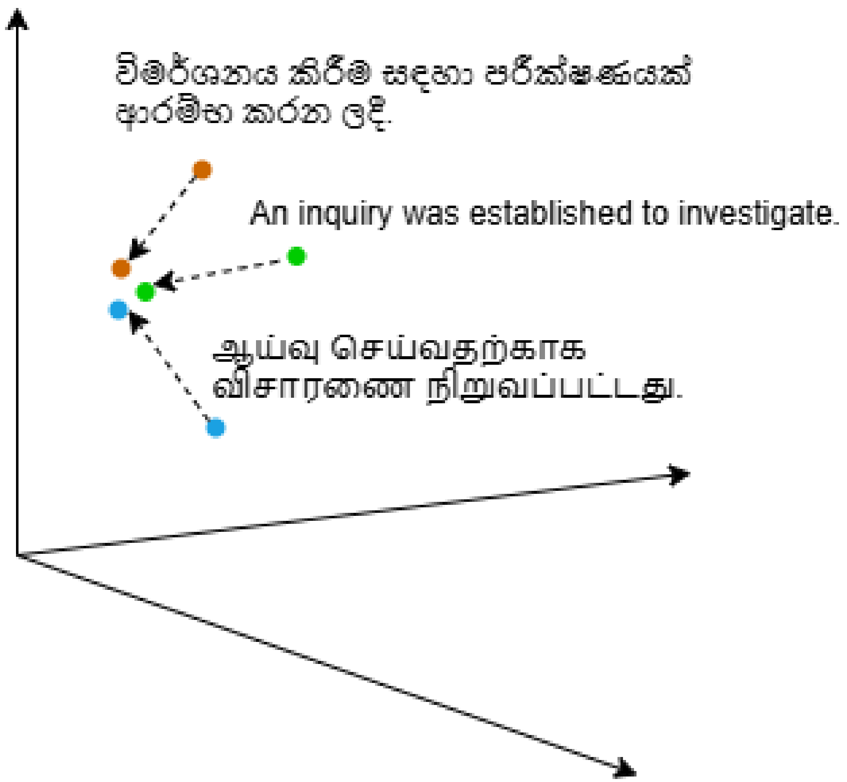
Fernando, A., Ranathunga, S., Sachintha, D., Piyarathna, L., Rajitha, C. (2023). Exploiting bilingual lexicons to improve multilingual embedding-based document and sentence alignment for low-resource languages. Knowledge and Information Systems, 65(2), 571-612. **(Know. And Info. Systems, 2024) Qartile: Q2; h-Index: 100**

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

Improving Representations in multiPLMs with Linguistic Entity Masking (LEM)



RO3 : Motivation



- Under-representation of monolingual training data during the pre-training stage. (Feng et al., 2022)
- Lack of explicit training objective to improve cross-lingual embedding. (Hu et al., 2020)

RO3 : Literature Review

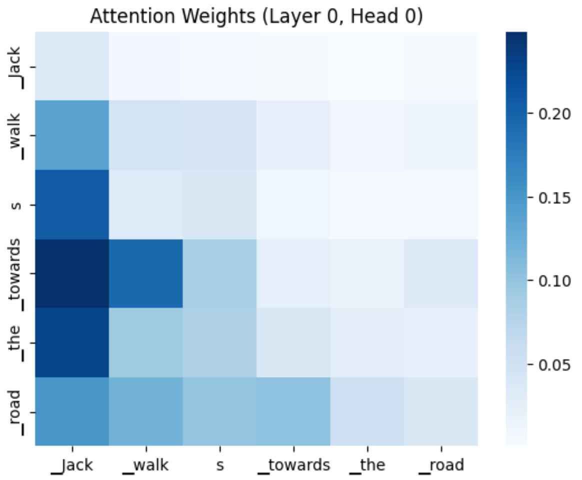
Encoder-based multiPLM models such as mBERT (Devlin et al., 2019), XLM-R (Conneu et al., 2020) are trained using Masked Language Modelling (MLM) objective to learn multilingual embeddings.

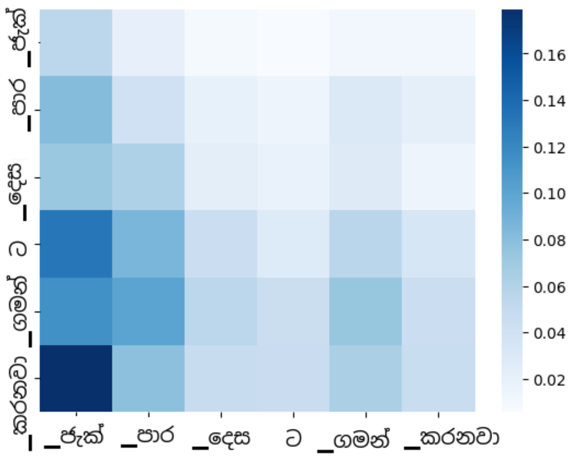
Masking Strategies	Monolingual/ Parallel	Sentence-retrieval Task Evaluation	Languages
Sub-word masking (Devlin et al., 2019)	Mono	✗	15 Languages
whole-word masking (Devlin et al., 2019)	Mono	✗	English
Entity/Phrase masking (Sun et al., 2019)	Mono	✗	English/Chinese
span-masking (Joshi et al., 2020)	Mono	✗	English
Point-wise Mutual Information-masking (PMI) (Levine et al., 2020)	Mono	✗	English
Translation Language Modelling (TLM) (Lample and Connaue, 2020)	Mono + Para	✓	15 Languages

Hypothesis : The cross-lingual alignment in existing multiPLMs can be improved with parallel data in a continual pre-training step

RO3: Assumptions / Methodology Decisions

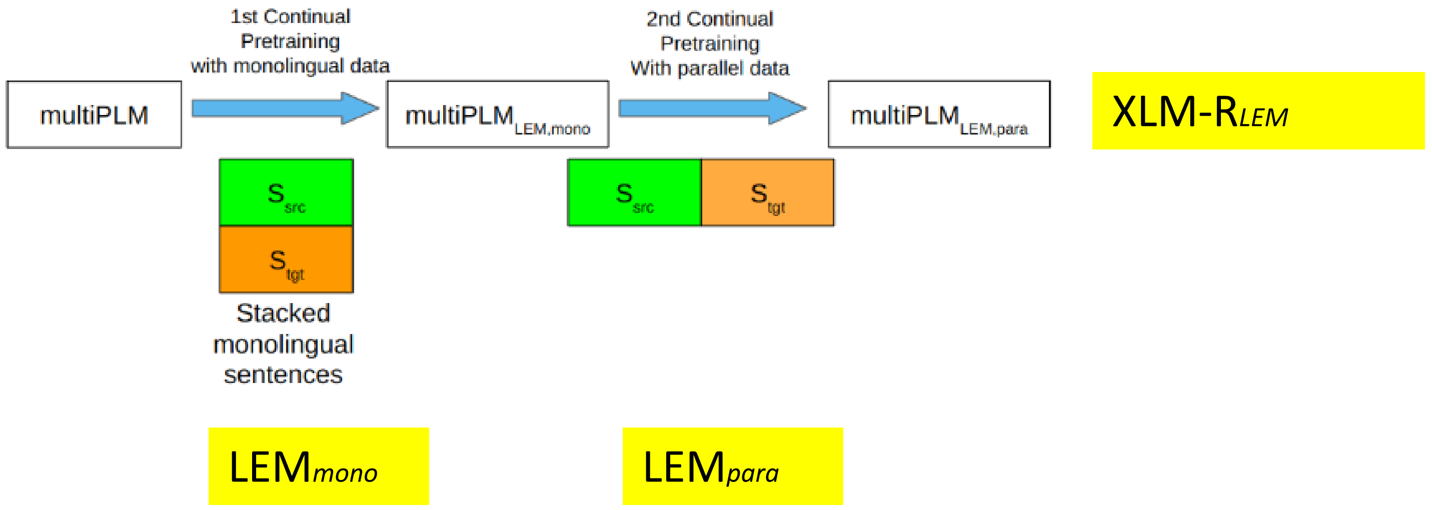
1. What are Linguistic Entities ?
    - Named Entities, Nouns and Verbs
2. Why masking Linguistic Entities?
    - Named Entities, Nouns and Verbs contribute to defining the syntactic and semantic structure of a sentence (Tenny et al., 2019)



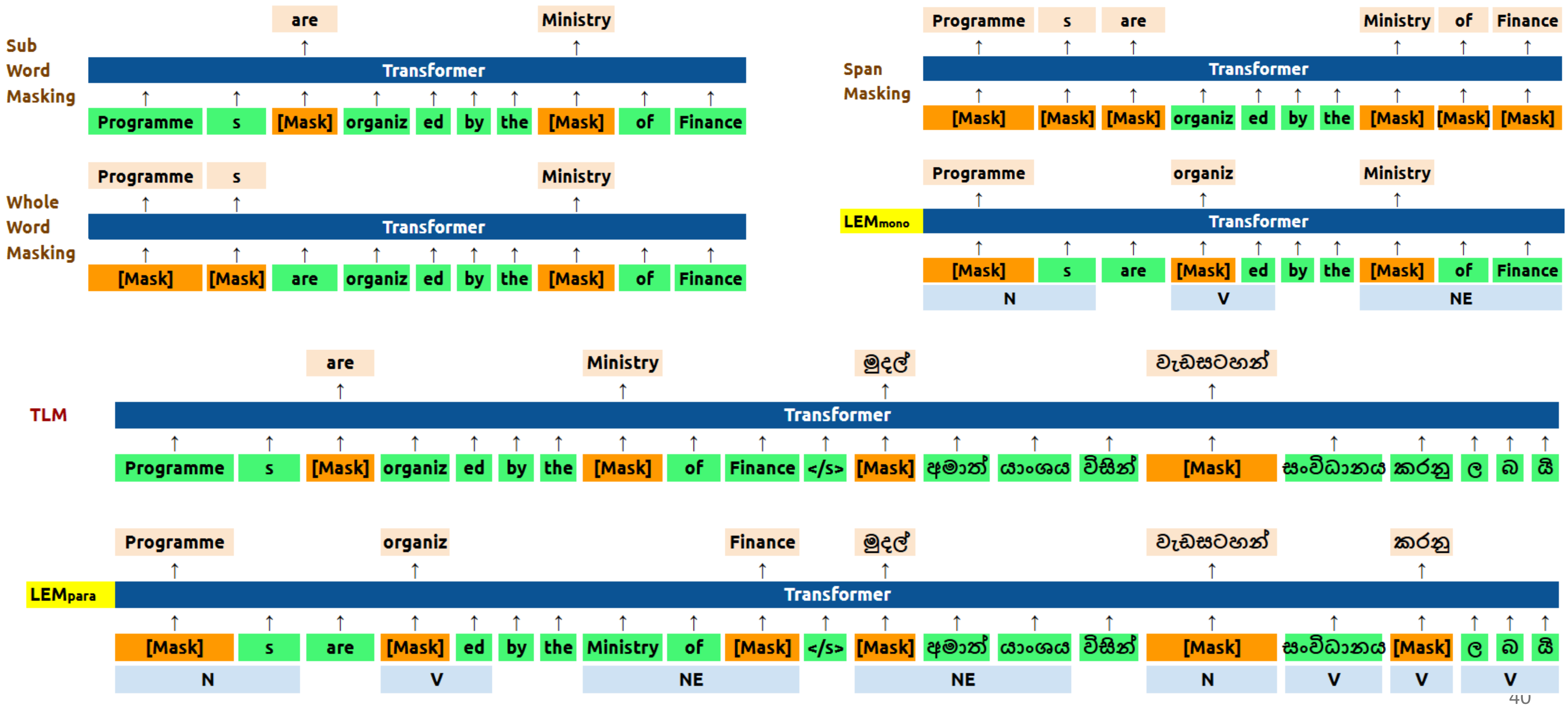


RO3: Assumptions / Methodology Decisions

3. Why in a continual pre-training step?



RO3 : Methodology



RO3: Methodology

Linguistic Entity Masking (*LEM_{mono}*)

$X = x_1\ x_2\ x_3\ ...x_i...\ x_n$ where x_i is a word and n is the number of words in the sequence.

After tokenization: $\bar{X} = \bar{x}_1\ \bar{x}_2\ \bar{x}_3\ \bar{x}_4.....\ \bar{x}_j....\ \bar{x}_m$

Identify Linguistic Entities: $\bar{X} = \{\{\bar{x}_1\ \bar{x}_2\}, ...\ \{\bar{x}_4\bar{x}_5\bar{x}_6\}, ...\{\bar{x}_m\}\}$

15% of tokens are masked from the sequence.

Continual pre-training objective: $$\mathcal{L}_{LEM_{mono}} = -\frac{1}{N}\sum_{j=1}^{N} y_j \log(P(x_j))$$

RO3: Methodology

Linguistic Entity Masking (*LEM$_{para}$*)

Concatenated parallel sentence pair : $\bar{Z} = \bar{x}_1 \ \bar{x}_2 \ \bar{x}_3 \ldots\ldots \ \bar{x}_k \ \bar{y}_1 \ \bar{y}_2 \ \bar{y}_3 \ldots\ldots \ \bar{y}_l$

15% of tokens are masked from the concatenated parallel sentence.

Continual pre-training objective :

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

RO3: Experiments

- Language pairs- En-Si, En-Ta and Si-Ta
- Our initial multiPLM is the XLM-R[1] base model.
- Evaluation
  - Intrinsic Evaluation – **Primary Task is sentence alignment**. We use Gold standard sentence alignment dataset [3]

[1] https://huggingface.co/FacebookAI/xlm-roberta-base
[2] https://github.com/UKPLab/sentence-transformers
[3] https://huggingface.co/datasets/NLPC-UOM/sentence_alignment_dataset-Sinhala-Tamil-English
[4] https://huggingface.co/datasets/allenai/nllb

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.

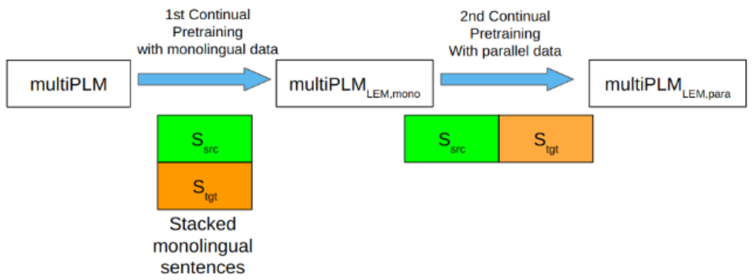
RO3: Experiments & Results

1. Empirical study of different masking strategies on the sentence alignment task.

Table 4 Bitext mining Recall scores for the different masking strategies

Experiment	Army			Hiru			ITN			Newsfirst			Averages		
	F	B	I	F	B	I	F	B	I	F	B	I	F	B	I
Sinhala - English															
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
Tamil - English															
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

Existing masking strategies shows reduced results compared to XLM-R baseline. Hence not favourable for improving cross-lingual representations.



RO3: Experiments & Results

2. Effectiveness of monolingual sides from the parallel data during the LEM_{mono} step.

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

Selecting monolingual sides from a parallel dataset improves performance in LEM$_{mono}$ step.

RO3: Experiments & Results

3. LEM Ablation experiments to identify the impactful linguistic entity for LEM$_{mono}$ and LEM$_{para}$: En-Si

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	93.33	91.00	93.92	93.67	90.83
100% NN+15% TLM on (100%NE+100%VB+100%NN+15%MLM)	89.67	91.33	86.00	95.02	95.35	93.33	93.67	94.67	91.33	93.67	94.00	90.33	93.00	93.84	90.25
100% NE+ 100%VB+15% TLM on (100%NE+100%VB+100%NN+15%MLM)	88.67	92.00	85.33	93.69	95.68	92.69	93.00	95.67	90.33	95.00	95.33	92.33	92.59	94.67	90.17
100% NE+ 100%NN+15% TLM on (100%NE+100%VB+100%NN+15%MLM)	86.67	91.67	84.67	96.35	96.01	94.68	94.33	95.67	92.33	94.33	94.67	91.33	92.92	94.50	90.75
100%NE+100%VB+100%NN+15%TLM on (100%NE+100%VB+100%NN+15%MLM)	91.33	92.00	88.00	95.68	95.68	94.02	93.67	94.00	91.33	93.67	94.00	90.33	93.59	93.92	90.92

RO3: Experiments & Results

3. LEM Ablation experiments to identify the impactful linguistic entity for LEM_{mono} and LEM_{para} : En-Ta

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

RO3: Experiments & Results

3. LEM Ablation experiments to identify the impactful linguistic entity for LEM_{mono} and LEM_{para} : SiTa

Experiment	Army FW	Army BW	Army IN	Hiru FW	Hiru BW	Hiru IN	ITN FW	ITN BW	ITN IN	Newsfirst FW	Newsfirst BW	Newsfirst IN	Average FW	Average BW	Average 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

RO3: Experiments & Results

3. LEM Ablation experiments during LEM_{mono} and LEM_{para} : Summary

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

- Compared to random token masking (Conneau and Lample, 2019) and **LEM strategy is effective** for cross-lingual representation improvement – across **language-pairs.**
- **Verbs** and **Named Entities** masked contributed to produce best gains.

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

RO3: Experiments & Results

4. **Secondary Sentence Retrieval Task** - Parallel data Filtration : ChrF++ scores for NMT

	ChfF++ Scores		
	Sinhala - Tamil	**English - Tamil**	**English - Sinhala**
XLM-R	33.58	38.28	30.37
MLM+TLM	35.98	45.35	39.78
$LEM_{mono}LEM_{para}$	**36.68**	**45.86**	**40.31**

5. **Fine-tuning Task** -Sentiment Classification for Code-mixed En-Si Dataset

En-Si Experiment	Precision	Recall	F1
XLM-R Ft Model	70.24%	**74.28%**	71.92%
XLM-R improved with TLM + MLM Ft Model	**75.35%**	69.24%	71.49%
XLM-R improved with LEMpara + LEMmono Ft Model	71.55%	72.72%	**72.11%**

- Results consistently show that LEM improved encoder is effective.

RO3: Limitations & Future Work

Limitations	Future Work
Performance of linguistic tools/models to identify NEs, Nouns and Verbs can limit the improvement. (False Positives/False Negative Examples)	Extend this study unified to train a single improved encoder catering several language-pairs.
Produce and improved multiPLM for eachlanguage-pair.	Upon releasing improved NEs and POS Taggersfor Sinhala/Tamil we will re-evaluate the workfor improvement.

RO3: Contributions & Publication

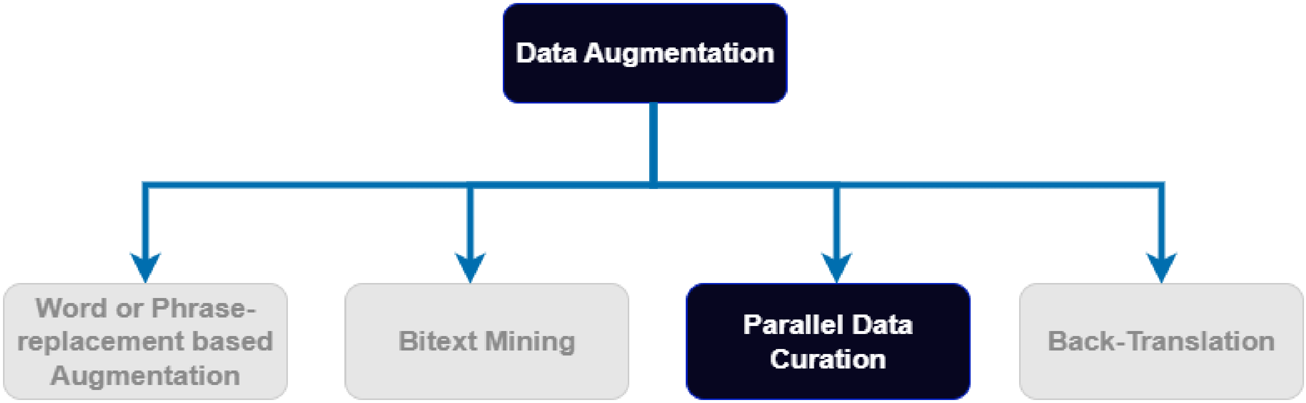
- Introduce an objective masking strategy termed Linguistic Entity Masking (LEM), to improve the cross-lingual representations of existing multiPLMs.
- This has been done using sentences from a parallel corpus with 56K only.
  Hence favourable for LRLs
- Publicly release the improved encoders for En-Si, En-Ta and Si-Ta language-pairs.

Publication

Fernando, A., Ranathunga, S. Linguistic entity masking to improve cross-lingual representation of multilingual language models for low-resource languages. Knowl Inf Syst (2025). https://doi.org/10.1007/s10115-025-02520-4 **(Know. And Info. Systems, 2024) Qartile: Q2; h-Index: 100**

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

Debiasing the Disparity in NMT systems



RO4: Motivation

- Web mined corpora is available for LRLs eg: **CCAligned** (El-Kishky et al., 2020), **CCMatrix** (Schwenk et al., 2021) and **ParaCrawl** (Bañón et al., 2020)
- Confirmed by Quality audits (Ranathunga et al., 2024;Kreutzer et al., 2022, Bane et al.,2022)
- NMT Models are sensitive to noise (Khayrallah and Koehn, 2018)
- PDC for LRLs has been emphasized with the introduction of WMT shared tasks (Sloto et al., 2023; Koehn et al., 2020, 2019)

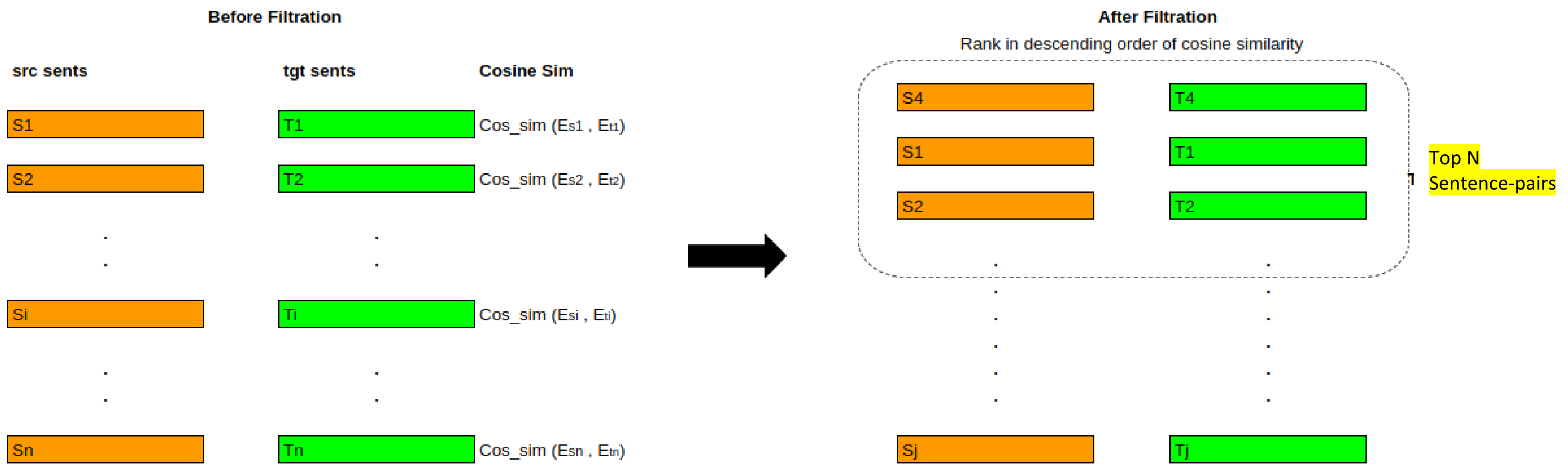
El-Kishky, A. and Guzmán, F. (2020). Massively multilingual document alignment with cross-lingual sentence-mover's distance. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing , pages 616–625, Suzhou, China. Association for Computational Linguistics.
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.
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.
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.
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.
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.
Khayrallah, H., & Koehn, P. (2018, July). On the Impact of Various Types of Noise on Neural Machine Translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation* (pp. 74-83).

RO4: What is Parallel Data Curation (PDC)?

Common approach for PDC : rank sentences according to the **semantic similarity (cosine similarity)** between sentence embeddings obtained for the source and target sentence pair.

The sentence representations (sentence embeddings) are obtained from a **multiPLM**

Select top N sentence-pairs and train a NMT system (Koehn et al., 2019; Koehn et al., 2020; Sloto et al., 2023)



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., 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.
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.

RO4: Motivation

Existing work has reported using different multiPLMs for ranking result in a disparity among the NMT scores. (Ranathunga et al., 2024, Moon el al., 2023)

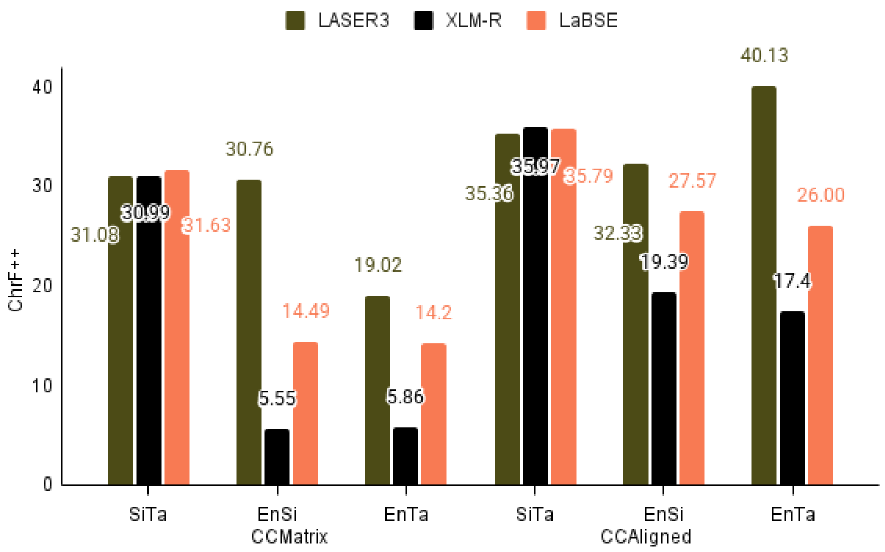
Three language-pairs

En-Si, En-Ta and Si-Ta

Corpus

CCMatrix , CCAligned

NMT models trained with top 100K ranked with
LASER3 (Heffernan et al., 2022)
XLM-R (Conneau et al., 2020) and
LaBSE (Feng et al., 2022)



Conducted human evaluation to analyse what type of sentences ranked by each multiPLM as high

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

RO4: Motivation – Human Evaluation

	CC	CN	CB	C	CS	CCN	UN	X	WL	NL	E
Sinhala - Tamil											
CCMatrix											
LASER3-Before	8%	27%	2%	**37%**	14%	14%	34%	1%	0%	0%	**63%**
XLM-R-Before	1%	10%	0%	**11%**	40%	19%	29%	0%	1%	0%	**89%**
LaBSE - Before	4%	6%	0%	**10%**	74%	7%	9%	0%	0%	0%	**90%**
CCAligned											
LASER3-Before	3%	24%	3%	**30%**	34%	19%	17%	0%	0%	0%	**70%**
XLM-R-Before	0%	0%	2%	**2%**	48%	49%	0%	0%	0%	1%	**98%**
LaBSE - Before	0%	1%	0%	**1%**	69%	26%	3%	0%	0%	1%	**99%**

	CC	CN	CB	C	CS	CCN	UN	X	WL	NL	E
English - Sinhala											
CCMatrix											
LASER3-Before	17%	7%	4%	**28%**	7%	10%	55%	0%	0%	0%	**72%**
XLM-R-Before	1%	0%	0%	**1%**	13%	4%	80%	2%	0%	0%	**99%**
LaBSE - Before	13%	2%	0%	**15%**	63%	14%	8%	0%	0%	0%	**85%**
CCAligned											
LASER3-Before	2%	22%	8%	**32%**	13%	30%	23%	2%	0%	0%	**68%**
XLM-R-Before	2%	0%	0%	**2%**	72%	20%	6%	0%	0%	0%	**98%**
LaBSE - Before	0%	1%	0%	**1%**	97%	2%	0%	0%	0%	0%	**99%**

	CC	CN	CB	C	CS	CCN	UN	X	WL	NL	E
English - Tamil											
CCMatrix											
LASER3-Before	0%	3%	2%	**5%**	0%	0%	95%	0%	0%	0%	**95%**
XLM-R-Before	0%	0%	2%	**2%**	3%	5%	90%	0%	0%	0%	**98%**
LaBSE - Before	0%	9%	2%	**11%**	34%	7%	48%	0%	0%	0%	**89%**
CCAligned											
LASER3-Before	2%	23%	18%	**43%**	13%	27%	17%	0%	0%	0%	**57%**
XLM-R-Before	0%	8%	4%	**12%**	42%	16%	15%	8%	0%	7%	**88%**
LaBSE - Before	0%	1%	0%	**1%**	97%	0%	0%	0%	0%	2%	**99%**

RO4: Motivation – Human Evaluation

Sinhala - Tamil	CC	CN	CB	C	CS	CCN	UN	X	WL	NL	E
CCMatrix											
LASER3-Before	8%	27%	2%	**37%**	14%	14%	34%	1%	0%	0%	**63%**
XLM-R-Before	1%	10%	0%	**11%**	40%	19%	29%	0%	1%	0%	**89%**
LaBSE - Before	4%	6%	0%	**10%**	74%	7%	9%	0%	0%	0%	**90%**
CCAligned											
LASER3-Before	3%	24%	3%	**30%**	34%	19%	17%	0%	0%	0%	**70%**
XLM-R-Before	0%	0%	2%	**2%**	48%	49%	0%	0%	0%	1%	**98%**
LaBSE - Before	0%	1%	0%	**1%**	69%	26%	3%	0%	0%	1%	**99%**

English - Sinhala	CC	CN	CB	C	CS	CCN	UN	X	WL	NL	E
CCMatrix											
LASER3-Before	17%	7%	4%	**28%**	7%	10%	55%	0%	0%	0%	**72%**
XLM-R-Before	1%	0%	0%	**1%**	13%	4%	80%	2%	0%	0%	**99%**
LaBSE - Before	13%	2%	0%	**15%**	63%	14%	8%	0%	0%	0%	**85%**
CCAligned											
LASER3-Before	2%	22%	8%	**32%**	13%	30%	23%	2%	0%	0%	**68%**
XLM-R-Before	2%	0%	0%	**2%**	72%	20%	6%	0%	0%	0%	**98%**
LaBSE - Before	0%	1%	0%	**1%**	97%	2%	0%	0%	0%	0%	**99%**

English - Tamil	CC	CN	CB	C	CS	CCN	UN	X	WL	NL	E
CCMatrix											
LASER3-Before	0%	3%	2%	**5%**	0%	0%	95%	0%	0%	0%	**95%**
XLM-R-Before	0%	0%	2%	**2%**	3%	5%	90%	0%	0%	0%	**98%**
LaBSE - Before	0%	9%	2%	**11%**	34%	7%	48%	0%	0%	0%	**89%**
CCAligned											
LASER3-Before	2%	23%	18%	**43%**	13%	27%	17%	0%	0%	0%	**57%**
XLM-R-Before	0%	8%	4%	**12%**	42%	16%	15%	8%	0%	7%	**88%**
LaBSE - Before	0%	1%	0%	**1%**	97%	0%	0%	0%	0%	2%	**99%**

MultiPLM –bias

Short Sentence (CS): Correct Translation, but the number of tokens on the Source or Target side is less

En - Si	LaBSE	Account Number	ගිණුම් අංකය
En - Si	LaBSE	11 July 2015.	11 ජූලි 2015.
En - Ta	XLM-R	July 21:	ஜூலை 21:
Si - Ta	XLM-R	ගණනය: 40 / 2, 40 / 3, 30 ආදිය.	எண்: 40 / 2, 40, 3, 30 முதலியன

Untranslated Text (UN): either in source or target side just copied from the translation counterpart

En - Si	XLM-R	What do you mean when you say "Your comment is awaiting moderation?"	මොකෝ විචාරක තුමා මගේ කමෙන්ට් එක තාම" Your comment is awaiting moderation
En - Ta	XLM-R	Effective Pixels: 16.0 million (Image processing may reduce the number of effective pixels.)	ஆப்டிகல் சென்சார் ரெசொலூஷன் 20.1 million (Image processing may reduce the number of effective pixels)

Overlapping Untranslatable Content (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

Hypothesis : Rule-based heuristics can be used to remove some of these noisy sentences.

RO4: Related Work – using multiPLMs in PDC

- **WMT2023 Shared Task uses LASER2 for ranking sentence-pairs (**Sloto et al., 2023)
- Gala et al. (2023) uses LaBSE during filtration of noise to train the NMT models
- Studies by Ranathunga et al. (2024) and Moon et al (2023) report disparity among different NMT models trained using ranked parallel corpors using multiPLMs.

No systematic study to identify the biases with respective to the multiPLMs in the top ranked parallel corpora.

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.
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.
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 .
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.
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 theAssociation for Computational Linguistics (Volume 1: Long Papers), pages 860–880.

RO4: Related Work on Heuristics used in PDC

- **Deduplication**
  - Remove identical duplicates (Costa-jussa et al., 2022)
  - Deduplicate after removing non-alpha characters and punctuations (Bala Das et al., 2023)
- **Length-based**
  - Removing short sentences (Gala et al., 2023; Aulamo et al., 2023)
  - Short sentences hinder NMT in two ways firstly, they have insufficient syntactic and semantic information secondly or can lead to overfitting (Koehn and Knowles, 2017)
- **LID- based**
  - Sentences with partial/full translations is a hindrance for learning seq-to-seq mappings
- **Ratio-based**
  - Can remove sentence-pairs with structural imbalances. Eg: source-to-target length ratio (Rossenbach et al., 2018; Gale and Church, 1993), alpha words-to-sentence length Ratio (Aulamo et al., 2020), alpha characters-to-sentence character ratio (Hangya and Fraser, 2018)

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 .
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.
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 machinetranslation models for all 22 scheduled indian languages. arXiv preprint arXiv:2305.16307 .
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.
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.
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.
Gale, W. A. and Church, K. (1993). A program for aligning sentences in bilingual corpora. Computational Linguistics , 19(1):75–102.
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.

RO4: Methodology – Parallel Sentences Categorization Taxonomy

Improvements to the existing parallel Sentences Categorization Taxonomy (Ranathunga et al., 2024)

Final Taxonomy	Description	Revision	Ranathunga et al., 2024
Quality Classes			
CC	Perfect Translation pair		Same
CN	Near Perfect Translation Pair		Same
CB	Weak Translation pair	We included over/under translations on the source or target to be included into the same category	
Noisy Classes			
CCN	Number/acronym/URL/email overlaps	The perfect or near perfect translaion pairs with more than 30% of the overlapping content is numbers/acronymns/URLs/email addresses (which cannot be translated/ transliterated)	New
CS	Short Sentences (Max 3 words)	Less than 5 words on either side	Modified
UN	Untranslated/Copied Text from source/target	Specifically define untranslated text as content which could have been translated/transliterated.	Modified
X	Mis-aligned sentence pair		Same
WL	Wrong Language (source/target)	To distinguish between UN defined an acceptable threshold as 30% for source/target	Modified
NL	Non-Linguistic (source/target)		Same

Examples for 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

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.

RO4: Methodology – Heuristic Selection

Mapping between the Noise category and the Heuristic Classes

Short Label	Noise Category	Heuristic Class
NL	Non-Linguistic	LID, sentWRatio/sentCRatio
WL	Wrong Language	LID
UN	Untranslated	LID
CS	Short Sentences	sLength
CCN	Number/acronym/URL/email overlaps	LID, sentWRatio/sentCRatio
X	Wrong Translations	**STRatio (with length difference)
CB	Weak Translations - Over/Under Translations	STRatio

RO4: Experiments

- Conduct the study across language pairs En-Si, En-Ta and Si-Ta
- Use two web-mined parallel corpora CCMatrix (Artetxe and Schwenk, 2019b) and CCAligned (El-Kishky et al., 2020)
- MultiPLMs – **LASER3, XLM-R and LaBSE** proven for cross-lingual tasks.
- Conduct Ablation studies
  - Find most impactful **individual** heuristic
  - Find optimal **heuristic combination**
- Evaluate the impact of the heuristic-based PDC on the disparity among NMT models
- Conduct Human evaluation to quantify the noise after heuristic based filtration

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

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.
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.
Imani, A., Lin, P., Kargaran, A. H., Severini, S., Sabet, M. J., Kassner, N., ... & Schütze, H. (2023, July). Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1082-1117).
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.

RO4 : Heuristic based PDC Experiments

- Each heuristic applied to Source (S), Target (T) and both sides (ST)
- **Deduplication** – Different granularities of de-duplication
  Identical de-duplication (dedup)
  de-duplication by removing numbers (nums)
  De-duplication by removing both numbers and punctuations (punctNums)
  N-gram deduplication (ngrams) where n=4,5,6,7
- **Length-based** (sLength = 3,4,5)
- **LID-based** : LID and LID with Threshold 0.7
- **Ratio-based** : STRatio 0.79-1.39 (EnSi), 0.87-1.62 (EnTa) and 0.85-1.57 (SiTa) were selected as thresholds for En-Si, En-Ta and Si-Ta respectively. ***sentWRatio*** and ***sentCRatio*** were taken as 0.6.

Finally, we train vanilla transformer-based NMT models with top 100k sentences from eachcorpus (CCMatrix, CCAligned), each Language pair (EnSi, EnTa, SiTa). Report NMT score usingChrF++ (Popovi et al., 2017).

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

RO4 : Experiments and Results.

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	

RO4 : Experiments and Results.

1. **De-duplication Ablation Experiments**
   - De-duplicating both Source and Target (94%) out perform de-duplicating Either Source (89%) or Target (83%) sides.
   - Conducting *dedup+ngram* (n=4,5,6,7) produced best result compared to *dedup*. Mostly it was **n=5,6.** However n is dependent on corpus characteristics.
   - ***dedup+punctNums*** outperforms **d*edup+nums*** or ***dedup***. (*dedup+punctNums vs* d*edup+nums* 78% and *dedup+punctNums vs* d*edup 67%)*
2. **sentLength**
   - *SLength = 5* as optimal sentence for filtration
   - SentLength filtrering both Source and Target is effective 56% of the time.
3. **LID-based Heuristics**
   - LID wih Threshold outperforms LID in 72% of the times. Therefore LID with Threshold recommended
4. **Ratio-based Heuristics**
   - Out of the three, (STRatio, sentWRatio and sentCRatio), **sentWRatio** performs best compared to its counterparts in 67% of experiments

RO4 : Experiments and Results - Summary

1. **What is the best performing individual heuristic?**
   - LIDThreshold 44%, de-dup experiments 33% and sLength 17%. Therefore the individual heuristic is dependent on the corpus characteristics

2. **Impact of the combined heuristics on the NMT results**
   - Highest NMT scores observed for combination except with CCMatix-SiTa language pair. Filtration produced less than 100k sentences.
   - ***dedup+punctNum+(n)gram+sLength+LIDThresh*** performed best while the ratio-based heuristic varied.
   - Exception CCAligned-SiTa performed best without LIDThresh. However results were compared to above combination (lags by –0.19 ChrF++).
   - ***dedup+punctNum+(n)gram+sLength+LIDThresh+sentWRatio*** produced best gains in 80%

RO4 : Experiments and Results - Summary

3. Impact of the heuristics on the disparity

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

LASER3 produce highest baseline NMT score.

Baseline disparity (Δ) = $\text{baseline}_{\text{LASER3}} - \text{baseline}_{\text{LM-R/LaBSE}}$

Disparity (%) after each individual/combined heuristic

$$\text{Disparity Reduction (\%)} = \frac{\Delta_{\text{baseline}} - \Delta_{\text{heuristic}}}{\Delta_{\text{baseline}}}$$

- Disparity among NMT scores in XLM-R-vs-LASER3 and LaBSE-vs-LASER3 reduced drastically with combined heuristics
- CCMatrix-EnSi and CCMatrix-EnTa reduction is around 50%. Which means there's still noise in the top ranked corpus.

RO4 : Human Evaluation

	CC	CN	CB	C	CS	CCN	UN	X	WL	NL	E
Sinhala - Tamil											
CCMatrix											
LASER3-Before	8%	27%	2%	**37%**	14%	14%	34%	1%	0%	0%	**63%**
LASER3-After	16%	68%	1%	**85%**	1%	4%	10%	0%	0%	0%	**27%**
XLM-R-Before	1%	10%	0%	**11%**	40%	19%	29%	0%	1%	0%	**89%**
XLM-R - After	0%	32%	2%	**34%**	1%	29%	35%	0%	1%	0%	**78%**
LaBSE - Before	4%	6%	0%	**10%**	74%	7%	9%	0%	0%	0%	**90%**
LaBSE - After	29%	33%	0%	**62%**	2%	32%	4%	0%	0%	0%	**38%**
CCAligned											
LASER3-Before	3%	24%	3%	**30%**	34%	19%	17%	0%	0%	0%	**70%**
LASER3-After	5%	79%	2%	**86%**	0%	9%	4%	1%	0%	0%	**14%**
XLM-R-Before	0%	0%	2%	**2%**	48%	49%	0%	0%	0%	1%	**98%**
XLM-R - After	20%	33%	4%	**57%**	1%	22%	19%	0%	1%	0%	**43%**
LaBSE - Before	0%	1%	0%	**1%**	69%	26%	3%	0%	0%	1%	**99%**
LaBSE - After	15%	34%	0%	**49%**	2%	43%	6%	0%	0%	0%	**51%**

	CC	CN	CB	C	CS	CCN	UN	X	WL	NL	E
English - Tamil											
CCMatrix											
LASER3-Before	0%	3%	2%	**5%**	0%	0%	95%	0%	0%	0%	**95%**
LASER3-After	6%	61%	20%	**87%**	0%	3%	10%	0%	0%	0%	**13%**
XLM-R-Before	0%	0%	2%	**2%**	3%	5%	90%	0%	0%	0%	**98%**
XLM-R - After	0%	39%	31%	**70%**	1%	3%	21%	4%	0%	1%	**30%**
LaBSE - Before	0%	9%	2%	**11%**	34%	7%	48%	0%	0%	0%	**89%**
LaBSE - After	36%	53%	4%	**93%**	1%	3%	2%	1%	0%	0%	**7%**
CCAligned											
LASER3-Before	2%	23%	18%	**43%**	13%	27%	17%	0%	0%	0%	**57%**
LASER3-After	3%	67%	10%	**80%**	0%	8%	12%	0%	0%	0%	**20%**
XLM-R-Before	0%	8%	4%	**12%**	42%	16%	15%	8%	0%	7%	**88%**
XLM-R - After	6%	46%	30%	**82%**	0%	9%	9%	0%	0%	0%	**18%**
LaBSE - Before	0%	1%	0%	**1%**	97%	0%	0%	0%	0%	2%	**99%**
LaBSE - After	19%	45%	3%	**67%**	0%	22%	11%	0%	0%	0%	**33%**

	CC	CN	CB	C	CS	CCN	UN	X	WL	NL	E
English - Sinhala											
CCMatrix											
LASER3-Before	17%	7%	4%	**28%**	7%	10%	55%	0%	0%	0%	**72%**
LASER3-After	39%	39%	7%	**85%**	0%	7%	8%	0%	0%	0%	**15%**
XLM-R-Before	1%	0%	0%	**1%**	13%	4%	80%	2%	0%	0%	**99%**
XLM-R - After	3%	8%	26%	**37%**	0%	2%	53%	8%	0%	0%	**63%**
LaBSE - Before	13%	2%	0%	**15%**	63%	14%	8%	0%	0%	0%	**85%**
LaBSE - After	87%	7%	3%	**97%**	0%	1%	2%	0%	0%	0%	**3%**
CCAligned											
LASER3-Before	2%	22%	8%	**32%**	13%	30%	23%	2%	0%	0%	**68%**
LASER3-After	13%	58%	14%	**85%**	0%	0%	13%	2%	0%	0%	**15%**
XLM-R-Before	2%	0%	0%	**2%**	72%	20%	6%	0%	0%	0%	**98%**
XLM-R - After	18%	18%	20%	**56%**	0%	6%	34%	4%	0%	0%	**44%**
LaBSE - Before	0%	1%	0%	**1%**	97%	2%	0%	0%	0%	0%	**99%**
LaBSE - After	45%	27%	3%	**75%**	1%	19%	5%	0%	0%	0%	**25%**

- **After applying heuristics qualitative improvement**
- Residual Noise CCN, UN which had not been filtered using heuristics

RO4: Limitations & Future Work

Limitations	Future Work
LID models sub-optimal performance for LRLs has an effect on filtration. Ie. UN sentences was to be removed from LID-based heuristic	Extend this work on how to eliminate specific noise categories, CNN and UN by means of training a classifier
We could not consider NLLB corpus into this analysis due to the computational limitations (NLLB EnSi-24M, EnTa-42M, SiTa-1.4M)	Heuristic filtration reduces 60% - 70% of dataset size. How best to use this filtered data into improving NMT results further(Steingrímsson et al., 2023)

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.

RO4: Contributions & Publication

- Empirically find heuristic combination leading to optimal NMT results and on the disparity among NMTmodels using multiPLM ranked parallel data.
- Improve existing taxonomy and conduct a comparative human evaluation to quantify noise before andafter heuristic-based filtration.
- Publicly release curated datasets CCMatrix and CCAligned for the three language-pairs

Publication

Fernando, A., Ranathunga, S., de Silva, N. Improving the quality of Web-mined Parallel Corpora of Low-Resource Languages using Debiasing Heuristics. arXiv preprint arXiv:2502.19074. **(Accepted. EMNLP 2025) Core Rank: A*/ h-Index: 193**

Conclusion

Contributions:

RO1. Propose and implement an algorithm to generate synthetic parallel sentences to augment OOV terms.

- Algorithm to generate synthetic parallel sentences by augmenting OOV terms, by imposing both syntactic and semantic features to validate.
- Publicly release the synthetic parallel sentences.

RO2: Empirical Study on the impact of the multiPLMs in the Document Alignment and Sentence Alignment tasks for LRLs.

- From empirical study, identifying that pre-trained models which had undergone continual pre-training with parallel data perform well for document alignment and sentence alignment tasks.
- Release the extended document alignment and sentence alignment evalution set, which was initially done by Rajitha et al., (2020)

Contributions:

RO3: Improving the cross-lingual representations of multiPLMs to identify High-Quality parallel sentences for the parallel sentence alignment task.

- Introduce an objective masking strategy termed Linguistic Entity Masking (LEM), to improve the cross-lingual representations of existing multiPLMs.
- This has been done using sentences from a parallel corpus with 56K only. Hence favourable for LRLs
- Publicly release the improved encoders for En-Si, En-Ta and Si-Ta language-pairs.

RO4 : Exploring parallel data filtration techniques to extract high-quality sentences from web-mined parallel corpora.

- Empirically find heuristic combination leading to optimal NMT results and on the disparity among NMT models using multiPLM ranked parallel data.
- Improve existing taxonomy and conduct a comparative human evaluation to quantify noise before and after heuristic-based filtration.
- Publicly release curated datasets CCMatrix and CCAligned for the three language-pairs

Publications:

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). **(PACLIC,2021) h5-Index: 13**

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. **(Know. And Info. Systems, 2024) Qartile: Q2; h-Index: 100**

Fernando, A., Ranathunga, S. Linguistic entity masking to improve cross-lingual representation of multilingual language models for low-resource languages. Knowl Inf Syst (2025). https://doi.org/10.1007/s10115-025-02520-4 **(Know. And Info. Systems, 2024) Qartile: Q2; h-Index: 100**

Fernando, A., Ranathunga, S., de Silva, N. Improving the quality of Web-mined Parallel Corpora of Low-Resource Languages using Debiasing Heuristics. arXiv preprint arXiv:2502.19074. **(Accepted. EMNLP 2025) Core Rank: A*/ h-Index: 193**

Other Publications

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

- 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). (**EACL, Best Paper Award Low Resource**)

- 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). **(ACL, 2024)**

- **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). **(ICON,2021)**
- **Fernando, A.**, Dias, G., & Ranathunga, S. (2021). Data augmentation and list integration for improving domain-specific Sinhala English-Tamil statistical machine translation.

Grants

- Awarded Travel Grant to present the poster “Linguistic Entity Masking to Improve Cross-Lingual Sentence Retrieval Capabilities of Multilingual Pre-trained Language Models for Low Resource Languages” at the 19th workshop for Women in Machine Learning (WiML) at the NeurIPS 2024 in Vancouver, Canada **(WiML workshop at NeurIPS, 2024)**
- Awarded Travel Grant to present the poster “Data Augmentation to Address the Out-of-Vocabulary Problem in Low-Resource Sinhala-English Neural Machine Translation” at the WiNLP: The Sixth Widening NLP Workshop at the Empirical Methods in Natural Language Processing (EMNLP) 2022 Conference **(WiNLP workshop at EMNLP, 2022)**

Acknowledgement

- National Languages Processing (NLP) Centre, at the University of Moratuwa for providing the GPUs to execute the experiments related to the research.

- 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.

- Senate Research Committee (SRC) grant of the University of Moratuwa, Sri Lanka, for funding the second part of the research.

- LK domain registry for funding me with the Prof. V. K. Samaranayake top-up grant during the third phase of this research.

- The Google Award for Inclusion Research (AIR) 2022 received by Dr Surangika Ranathunga and Dr Nisansa de Silva.

Thank you Note

- Supervisors Dr Nisansa de Silva, Dr Surangika Ranathunga
- Progress Review Panel, CSE Research Coordinator Dr Kutila Gunasekara
- The examination panel and the Chairman
- Prof. Gihan Dias, Dr Uthayasanker Thayasivam, current HoD and Prof. Sanath Jayasena
- All Academic and Non-Academic staff at the Department of Computer Science & Engineering, University of Moratuwa
- All the collogues at University of Moratuwa
- My parents, husband and children.

Refences

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., Çolako ˘glu, T., Hatipo ˘glu, 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 V ollgraf, 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., Stribi ˙zew, 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 58thAnnual 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. In29th Annual Meeting of the Association for Computational Linguistics , pages 169–176, Berkeley, California, USA. Association forComputational Linguistics.

Refences

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., Haji ˇc, 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 .
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 multistage fine-tuning for low-resource neural machine translation. In Proceedings of the 2019 Conference on Empirical Methods in Natural LanguageProcessing 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 languagetechnologies. 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. (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 theAssociation 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 theAssociation 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 LanguageResources 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.

Refences

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 , pages3799–3807.

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: ShortPapers) , 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 InternationalConference on Language Resources and Evaluation (LREC 2018) .

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: LongPapers) , 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 vocabulary problem 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 NaturalLanguage 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 all22 scheduled indian languages. arXiv preprint arXiv:2305.16307 .

Refences

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.
InProceedings 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., Ceau¸ su, 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.

Refences

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. InProceedings 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 ranslation: 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.

Refences

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.
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. InMachine 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. InProceedings 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.
Medved’, M., Jakubí ˇcek, M., and Ková ˇr, 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.

Refences

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 theEighth 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. InColing 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. InCOLING 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. InProceedings 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.
Popovi ´c, 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.
Popovi ´c, 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.

Refences

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 inNatural 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 EmergingRegions (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 InternationalConference on Asian Language Processing (IALP) , pages 244–249. IEEE.
Prószéky, G. (2005). An approach to machine translation via the rule-to-rule hypothesis. InProceedings 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). Adapterbased fine-tuning of pre-trained multilingual language models for code-mixed and code-switched text classification. Knowledge and Information Systems , 64(7):1937–1966.
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. InProceedings 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.

Refences

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.

Refences

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. In2017 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, 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.
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.
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., ¸ Sahın, 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.

Refences

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 NaturalLanguage 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.