Linguistic Entity Masking to Improve Cross-Lingual Representations in Encoder-based LLMs

Aloka Fernando[1], Surangika Ranathunga [2], Nisansa de Silva[1]
{alokaf,NisansaDdS}@cse.mrt.ac.lk
s.ranathunga@massey.ac.nz
[1]Dept. of Computer Science and Engineering, University of Moratuwa,10400, Sri Lanka
[2]Massey University, Palmerston North, New Zealand, 4443
Correspondence: alokaf@cse.mrt.ac.lk

Abstract

Multilingual Pre-trained Language Models (multiPLMs) trained with the Masked Language Modeling (MLM) objective exhibit suboptimal performance on cross-lingual downstream tasks for Low-Resource Languages (LRLs). Continually pre-training these multiPLMs with the Translation Language Modeling (TLM) objective on parallel data improves the cross-lingual performance. However, both MLM and TLM mask tokens randomly, which does not guarantee optimal representation learning. In this paper, we introduce a novel masking strategy, Linguistic Entity Masking (LEM) to improve the cross-lingual representations of existing multiPLMs. In contrast to MLM and TLM, LEM limits masking to the linguistic entities nouns, verbs and Named Entities, which hold a higher prominence in a sentence. We hypothesise that masking linguistically significant linguistic entities should contribute to effective representation learning. Empirically we prove this using two downstream tasks with three LRL pairs, English-Sinhala, English-Tamil, and Sinhala-Tamil, and show that our LEM-based learning returns superior results compared to MLM+TLM.

1 Introduction

Encoder-based Multilingual Pre-trained Language models (multiPLMs), trained on the Masked Language Modeling (MLM) objective (Devlin et al., 2019; Conneau et al., 2020), perform well for cross-lingual NLP downstream tasks (Conneau et al., 2020). However, for the same tasks, these multiPLMs return suboptimal results (Ruder et al., 2021; Hu et al., 2021) for Low-Resource Languages (LRLs). Feng et al. (2022) and Conneau and Lample (2019) conducted a further pre-training of such MLM pre-trained multiPLMs using a contrastive objective and a Translation Language Modeling (TLM) objective, respectively. Their findings showed that the cross-lingual capability of these multiPLMs can be enhanced in a continual pre-training step using parallel data. However, MLM and TLM select random tokens for masking. We believe that randomness is not objective enough for effective representation learning, and in Low-resource conditions, this can even degrade.

In the direction of masking-based representation learning, variants of the MLM strategy, such as Whole-word Masking (Devlin et al., 2019), Span Masking (Joshi et al., 2020), Point-wise Mutual Information (PMI) Masking or NER/Noun Phrase Masking (Sun et al., 2019) have been proposed. However, such strategies have only been evaluated for High-Resource Languages (HRLs), and have not been evaluated for their effectiveness on sentence retrieval tasks.

In contrast to existing work, we hypothesise that masking tokens from linguistically prominent Named Entities, Nouns, and Verbs would lead to effective representation learning and propose a Linguistic Entity Masking (LEM). The intuition is that for Pre-trained Language Models (PLMs) to capture the notion of syntactic structures and grammatical properties in the language (Nastase and Merlo, 2024, 2023; Aoyama and Schneider, 2022), these linguistic entities (ie. NEs, Verbs and Nouns) contribute significantly.

We improve existing encoder-based multiPLMs by applying the LEM strategy in a two-step continual pre-training process. First by applying LEM with monolingual data and secondly, with parallel data. We empirically prove that the cross-lingual representation improvement leads to superior results, compared to MLM+TLM improved encoder using two sentence retrieval tasks: Sentence alignment and Parallel Data Curation (PDC). We experimented this on three low-resource language pairs, English-Sinhala (En-Si), English-Tamil (En-Ta) and Sinhala-Tamil (Si-Ta).

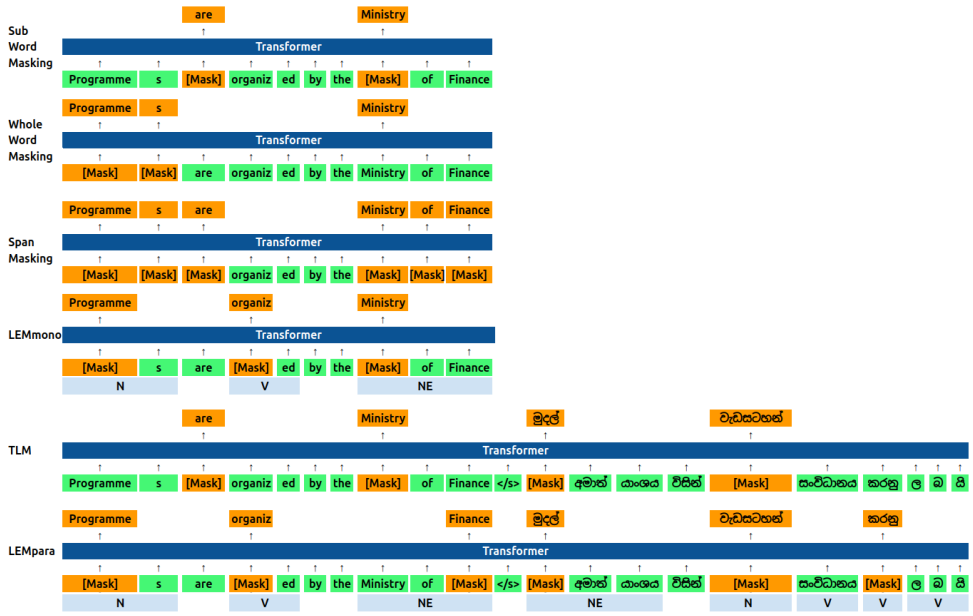

Figure 1: A comparison of the existing masking strategies considering an example from the En-Si language pair. Sub-word masking, Whole Word masking, span masking, and LEM_{mono} consider only monolingual sentences during masking. TLM and LEM_{para} consider concatenated parallel sentences to apply the masking.

2 Linguistic Entity Masking (LEM)

Our LEM strategy, in comparison with the existing masking strategies is illustrated in Figure 1. In Figure 2, we show the training pipeline, where an existing MLM pre-trained multiPLM, is continually pre-trained using the LEM strategy firstly on monolingual data and secondly on parallel data.

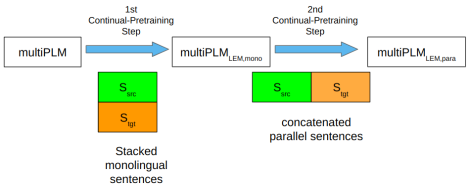

Figure 2: The LEM continual pre-training process. In the first step ie. LEM$_{mono}$ we continually pre-train an existing multiPLM using monolingual data. In the second continual pre-training step ie. LEM$_{para}$, the LEM strategy is applied on the *concatenated parallel data*.

.

Theoretical Framework

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

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

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

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

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

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

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

Finally, we extend the TLM objective with LEM into the parallel data setting (LEM$_{para}$). Here we concatenate the source sentence ($X = x_1\ x_2\ x_3\ .....\ x_m$) and target sentence ($Y = y_1\ y_2\ y_3\ .....\ y_n$) from the parallel sentence-pair as a single input example and obtain the tokenized

output as represented by $\bar{Z}$ in Equation 4. $\bar{X} = \bar{x}_1\ \bar{x}_2\ \bar{x}_3 \ldots\ldots\ \bar{x}_k$ and $\bar{Y} = \bar{y}_1\ \bar{y}_2\ \bar{y}_3 \ldots\ldots\ \bar{y}_l$ are the tokenized source and target sentences, respectively. k and l are the number of tokens (sub-words) in the source and target sentences (respectively) after tokenization.

$$\bar{Z} = \bar{x}_1\ \bar{x}_2\ \bar{x}_3 \ldots\ldots\ \bar{x}_k\ \bar{y}_1\ \bar{y}_2\ \bar{y}_3 \ldots\ldots\ \bar{y}_l \quad (4)$$

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

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

3 Experiments &Results

3.1 Data and multiPLM Selection

We experiment with three LRL-pairs, En-Si, En-Ta and Si-Ta. Since the languages belong to three different sub-families, we believe our more generalized. We use the SITA-Trilingual human-curated dataset (Fernando et al., 2020) during continual pre-training steps.

We select XLM-R (Conneau et al., 2020) as the base multiPLM for our experiments [1] as it has already demonstrated promising performance for the LRLs considered in this study (Dhananjaya et al., 2022; Rathnayake et al., 2022; Udawatta et al., 2024; Ranathunga et al., 2024).

3.2 Sentence Alignment Task

We use the sentence alignment task to primarily evaluate the masking strategies. Its performance depends on the quality of the cross-lingual embeddings, returned from the multiPLM. We follow the work of Artetxe and Schwenk (2019) and determine the target side sentence for each source side sentence by calculating the margin-based cosine similarity. This we call as the Forward (FW) criterion. In Backward (BW) criterion, for each target side sentence, a source side sentence is retrieved while the Intersection (IN) refers to the common sentence pairs retrieved with FW and BW criterion. This is detailed in Appendix A.

[1]Other popular multiPLMs, XLM and mBERT do not cover the Sinhala language.

3.3 Evaluation of Different Masking Strategies

First, we conduct an empirical evaluation of existing masking strategies and assess their performance on the sentence alignment task. The masking strategies explored in this study are shown in Table 1 while the detailed results are shown in Table 5 in Appendix B. Except for Si-Ta BW direction, for all other language pairs and sentence alignment criterion (FW, IN) the existing masking strategies performed less than the baseline.

Sinhala and Tamil are morphologically rich languages, where words are inflected based on number, gender and case. Further, these languages have more Out-of-Vocabulary (OOV) leading to a higher word-to-token ratios. Applying whole-word masking and span masking leads to longer token-spans, leading to limited context to predict the masked tokens accurately. We believe that this reduced context reduces the prediction accuracy, thus weakening the already learnt representations in the XLM-R model. As a result, the existing masking strategies return degraded results.

Masking Strategy	Averages		
	FW	BW	IN
Si-En			
XLM-R (Conneau et al., 2020)	**94.84**	**95.34**	**93.00**
Sub-word Masking (Devlin et al., 2019)	90.92	94.76	89.68
Whole-word Masking (Devlin et al., 2019)	92.25	93.00	90.34
Span Masking (Joshi et al., 2020)	92.00	92.01	88.26
Ta-En			
XLM-R (Conneau et al., 2020)	**86.31**	**85.39**	**80.71**
Sub-word Masking (Devlin et al., 2019)	84.64	83.55	78.55
Whole-word Masking (Devlin et al., 2019)	81.97	83.13	75.38
Span Masking (Joshi et al., 2020)	83.56	82.97	76.79
Si-Ta			
XLM-R (Conneau et al., 2020)	**89.78**	89.45	**86.20**
Sub-word Masking (Devlin et al., 2019)	89.60	**91.02**	86.07
Whole-word Masking (Devlin et al., 2019)	88.61	90.77	85.11
spanMasking (Joshi et al., 2020)	88.61	89.44	85.11

Table 1: Sentence alignment Recall scores for the different masking strategies.

3.4 Evaluation of LEM Strategy and Ablation Study

The objective of the LEM strategy is to improve cross-lingual representation of existing multiPLMs. Hence, we conduct an ablation study to identify the effectiveness of the LEM strategy, with respect to

each linguistic feature in isolation and in combinations. Thereby, we intend to identify the most impactful linguistic feature or feature combination for representation improvement. The Sentence alignment results are shown in Tables 10, 11, 12 in the Appendix C. In Table 2, we summarise the final gains.

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

Table 2: Results for Sentence alignment task in terms of recall points. For comparison purposes, the FW, BW and IN gains are averaged and reported in the *Overall Average Gain column*.

We identify the NEs using an in-house fine-tuned NER model (Ranathunga et al., 2024). We use a language-specific POS Tagger for Noun and Verb identification as in Appendix C. We customize the MLM training implementation released with the sentence-transformers[2] library (built on Huggingface transformers[3]), to support XLM-R tokenization and implement the LEM strategy. Further experimental details have been included in Appendix C.

It was observed that during the first continual training step itself, the LEM surpassed the XLM-R for En-Si and Si-Ta language pairs. For the En-Ta language pair, the LEM$_{mono}$ step produced a reduced score compared to the XLM-R baseline. However, after the second continual pre-training step, XLM-R$_{LEM}$ scores surpassed the XLM-R baseline by +1.2. Therefore, it is safe to say that LEM improves the cross-lingual representation for the language pairs.

Among the three language-pairs we observe that the LEM had been most effective for the En-Ta language pair, which was +2.4 recall points. Since English and Tamil belong to different language families, the cross-lingual representations in existing multiPLMs are weak. Therefore, the LEM$_{para}$ is effective in improving cross-lingual representations.

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

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

3.5 Parallel Data Curation

Finally, we evaluate the effectiveness of the improved encoder on the Parallel Data Curation (PDC) task for all three language pairs. Using the NLLB (Costa-jussà et al., 2022) and CCAligned (El-Kishky et al., 2020) datasets, we rank each sentence pair based on the cosine similarity between their embeddings, obtained from the multiPLMs. Then, using the top 50K sentence pairs, we train the NMT model between the language pairs and report results on the Flores+[4] devtest. Further details are in Appendix D. The NMT results are available in Table 3.

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

Table 3: ChrF scores for the parallel data curation task. The values in brackets indicate the gains of XLM-R$_{LEM}$ compared to the XLM-R and the XLM-R$_{MLM+TLM}$ respectively.

The NMT results show that the XLM-R$_{LEM}$ model produce superior results compared to XLM-R$_{MLM+TLM}$, for all three language pairs across the two corpora. We believe the magnitude of the gain is dependent on the characteristics of the parallel corpus and the size of the training data sample. For the English-Tamil language pair, the CCAligned corpus produce a significant gain for XLM-R$_{LEM}$ compared to XLM-R$_{MLM+TLM}$. This justifies the effectiveness of the LEM strategy which was not evident with random masking followed in MLM+TLM.

The rest of the gains vary from +0.3 to +0.8 ChrF points. According to metric analysis by Kocmi et al. (2024), these gains are equivalent to +0.48 to +1.12 BLEU points with a human accuracy of 54.2% to 66% respectively. This means the improvement in the translation quality in the NMT systems is almost in line with a minimum human accuracy rating of 54.2% to 66%. Therefore, improvement in the cross-lingual representations with the LEM strategy benefits the parallel data curation task as well.

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

4 Ablation Study : Number of Tokens for Masking in LEM Strategy

To evaluate the impact of the masked token count variation within linguistic entities, we conducted an ablation study by varying the number of masked tokens. We conducted this for the Sinhala-Tamil language pair. As reported in Table 12, the best result was returned for the 100% NE+15% MLM and 100% NE+15% TLM combinations in the LEM_{mono} and LEM_{para} steps respectively. Therefore, we used this setting and by varying the number of tokens for masking, we compared the impact on the bitext mining scores. The results are reported in Table 4. It reveals a clear trend of decreasing performance.

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

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

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

Table 4: Ablation study results by changing the number of tokens masked in the linguistic entity. The results are for the Sinhala-Tamil language pair and the bitext mining downstream task.

5 Conclusion

Continually pre-training a multiPLM with the TLM objective on parallel data improves cross-lingual representations of such models. However, both MLM and TLM select tokens for masking randomly, which may not guarantee optimal representation learning, particularly in a LR settings.

This research introduce LEM strategy, aimed at masking linguistic entities (NEs, nouns and verbs), in two continual pre-training steps using monolingual data and parallel data respectively. We select linguistic entities for masking as they significantly contribute to syntactic and semantics of the sentence. Empirically we prove that the LEM improved multiPLM outperforms the existing MLM+TLM improved representations for bitext-mining and parallel data curation tasks, confirming our hypothesis.

In contrast to the other masking strategies aimed at masking consecutive text spans, at the word-level, phrase-level or NE-levels, our ablation experiments show that single-token masking within the linguistic-entity is favourable for improving cross-lingual representations further. We believe that masking consecutive spans weakens the context for accurate token prediction, leading to sub-optimal representation learning.

Hence we conclude that the LEM strategy is favourable to improve the cross-lingual representations in a low-resource setting with only 56k parallel sentences.

Limitations

The final results of the LEM strategy are dependent on the accuracy of the NER model and POS Taggers. Although these perform well for En, we observe two main error types with it for Sinhala and Tamil, as shown in Table 8 in Appendix E. The primary limitation was that they return False positives and False negatives of these taggers. Firstly, False Positives, we observe words which were not a part of the NE tagged as NEs. Secondly, in the category of False Negatives, the NER model fails to identify all the words belonging to the NE sequence, and incorrectly label these words as Other etc. Similar instances are found in PoS Tagging as well as per examples in Table 9 in Appendix D that resulted in False Positives and False Negatives. However, for the English language, the returned PoS tags were mostly accurate.

Acknowledgments

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

This research was funded by the Google Award for Inclusion Research (AIR) 2022, received by Dr Surangika Ranathunga and Dr Nisansa de Silva.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Tatsuya Aoyama and Nathan Schneider. 2022. Probeless 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.

Mikel Artetxe and Holger Schwenk. 2019. 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.

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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

Vinura Dhananjaya, Piyumal Demotte, Surangika Ranathunga, and Sanath Jayasena. 2022. Bertifying sinhala-a comprehensive analysis of pre-trained language models for sinhala text classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7377–7385.

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

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

Aloka Fernando, Surangika Ranathunga, and Gihan Dias. 2020. Data augmentation and terminology integration for domain-specific sinhala-english-tamil statistical machine translation. *arXiv preprint arXiv:2011.02821*.

Aloka Fernando, Surangika Ranathunga, Dilan Sachintha, Lakmali Piyarathna, and Charith Rajitha. 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.

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

Sandareka Fernando, Surangika Ranathunga, Sanath Jayasena, and Gihan Dias. 2016. Comprehensive part-of-speech tag set and svm based pos tagger for sinhala. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 173–182.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3633–3643.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *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.

Vivi Nastase and Paola Merlo. 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.

Vivi Nastase and Paola Merlo. 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.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, pages 48–53.

Surangika Ranathunga, Asanka Ranasinghea, Janaka Shamala, Ayodya Dandeniyaa, Rashmi Galappaththia, and Malithi Samaraweeraa. 2024. A multi-way parallel named entity annotated corpus for english, tamil and sinhala. *arXiv preprint arXiv:2412.02056*.

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

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and 1 others. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245.

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

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Pasindu Udawatta, Indunil Udayangana, Chathulanka Gamage, Ravi Shekhar, and Surangika Ranathunga. 2024. Use of prompt-based learning for code-mixed and code-switched text classification. *World Wide Web*, 27(5):63.

A Sentence Alignment Task and Evaluation Set

Sentence alignment is a sentence retrieval task that retrieves a target language translation for a given source sentence or vice versa from a set of monolingual source-side and target side sentences. The performance of sentence alignment relies heavily on the quality of cross-lingual embeddings.

After obtaining sentence embeddings from the multiPLM, we use the margin-based cosine similarity (Artetxe and Schwenk, 2019) function to identify parallel sentence pairs. We choose margin-based cosine similarity over conventional cosine similarity for this task due to its lower rate of false positives. Then we rank the parallel sentences according to their similarity scores. Sentence alignment is performed using the three criteria: Forward (FW), Backward (BW), and Intersection (IN) following the work of (Artetxe and Schwenk, 2019). FW retrieves the target sentence for each source sentence, BW retrieves the source sentence for each target sentence, and IN considers the intersection of the parallel sentences retrieved using FW and BW criteria.

We evaluate the sentence alignment task, using the gold standard human-created evaluation set (Fernando et al., 2023). It consists of trilingual data obtained from four Sri Lankan news sources Army[5], Hiru[6], ITN[7] and Newsfirst[8]. For each news source, there are human-aligned 300 sentence-pairs. We report the results using Recall metric, as supported by this sentence alignment dataset.

B Empirical Evaluation on Existing Masking Strategies

We empirically evaluate the existing masking strategies and assess their performance on the sentence alignment task. The experimental results by means of Recall scores are shown in Table 5. The masking strategies explored in this study are as follows:

Sub-word Masking - Following the BERT MLM, with each sentence, 15% of tokens are selected randomly and corrupted according to 80%-10%-10% rule.
Whole Word Masking - All the sub-words corresponding to the randomly sampled words are masked. A total of 15% tokens are sampled and corrupted according to 80%-10%-10% rule.
Span Masking - Consecutive word spans are sampled over a geometrical distribution and 15% of tokens are masked. The masking is limited to whole-word tokens as defined in the original work.

C Linguistic Entity Masking Ablation Study

In this section, we present the full experiments along with the scores obtained during the ablation

[5]https://www.army.lk/
[6]https://www.hirunews.lk/
[7]https://www.itnnews.lk/
[8]https://english.newsfirst.lk/

Experiment	Army			Hiru			ITN			Newsfirst			Averages		
	FW	BW	IN	FW	BW	IN	FW	BW	IN	FW	BW	IN	FW	BW	IN
En - Si															
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
En - Ta															
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
Si-Ta															
XLM-R	83.44	81.46	78.15	**90.67**	91.00	**87.33**	**91.33**	90.00	87.00	**93.67**	**95.33**	**92.33**	**89.78**	89.45	**86.20**
Sub-word Masking	**86.75**	88.08	**81.96**	88.00	89.33	84.00	93.33	**92.67**	**89.33**	90.33	94.00	89.00	89.60	**91.02**	86.07
Whole-word Masking	85.76	**89.73**	81.46	88.33	**91.33**	84.67	90.33	90.33	86.67	90.00	91.67	87.67	88.61	90.77	85.11
spanMasking	85.78	85.10	81.79	88.67	91.00	87.00	91.00	91.00	87.33	89.00	90.67	84.33	88.61	89.44	85.11

Table 5: Sentence Alignment Recall scores for the different masking strategies.

study of our LEM masking strategy. The Tables 10, 11 and 12 contain the sentence alignment recall scores for En-Si, En-Ta and Si-Ta language pairs respectively.

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

To identify nouns and verbs in the sequences, we employ Flair POS tagger (Akbik et al., 2018) for English, the Sinhala TnT POS Tagger (Fernando and Ranathunga, 2018; Fernando et al., 2016) for Sinhala, and ThamizhiUDp (Sarveswaran and Dias, 2020) for Tamil. Then the checkpoint with the least validation loss is selected as the best-performing model. Each continual pre-training experiment is executed for 60 epochs with early stopping. The experiments are conducted on Nvidia Quadro RTX 6000 GPU with 24GB VRAM. The hyperparameters of XLM-R [9] model and other training parameters used in the continual pre-training experiments are shown in Table 6.

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

Table 6: Hyperparameters used during continual pre-training with the LEM strategy

D NMT Experiments

In this section, we outline the NMT experimental setup and hyperparameters used. First we train a Sentencepiece[10] tokenizer with a vocabulary size of 25000. Then we use fairseq toolkit (Ott et al., 2019) to model and train the vanilla transformer-based Sequence-to-Sequence NMT model. The experiments are conducted on a Nvidia Quadro RTX6000 GPU with 24GB VRAM. The hyperparameters used during training along with the training parameters are shown in Table 7. Each experiment is run for 100 epochs with the early stopping criteria.

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

Table 7: Training parameters for NMT experiments.

E Limitations of the NER Model and Pos Tagger

Examples highlighting the error categories found with the NER model and PoS taggers in the limitations section are shown in Table 8 and Table 9 respectively.

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

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

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

Table 8: Examples of incorrect identification and labelling of NEs. We identify mainly two error categories: false positives and false negatives, where the NER model underperforms.

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

Table 9: Examples of incorrect identification and labelling of PoS Tags. We identify mainly two error categories: false positives and false negatives, where the Pos Tagger underperforms.

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

Table 10: LEM Ablation experiments and sentence alignment scores for En-Si language pair considering linguistic entity masking. The best result for each criterion is shown in bold and highlighted in yellow.

Experiment	Army			Hiru			ITN			Newsfirst			Average		
	FW	BW	IN	FW	BW	IN	FW	BW	IN	FW	BW	IN	FW	BW	IN
Baselines															
XLM-R	86.67	88.33	82.00	83.00	78.33	72.67	83.22	83.56	78.86	92.33	91.33	89.33	86.31	85.39	80.71
15%MLM	84.00	86.00	77.67	80.33	75.00	68.33	81.56	82.21	78.52	90.67	91.00	87.00	84.14	83.55	77.88
15%TLM on 15%MLM	86.67	85.67	79.33	80.33	78.67	71.00	81.88	83.56	77.52	90.00	92.67	88.00	84.72	85.14	78.96
LEM$_{mono}$															
100% NE+15% MLM	86.00	86.67	81.00	79.33	75.33	66.67	81.21	81.21	74.83	93.00	92.00	90.00	84.89	83.80	78.12
100% VB+15% MLM	85.67	84.67	76.67	78.67	76.00	68.00	81.88	82.55	75.84	91.00	90.00	86.33	84.30	83.30	76.71
100% NN+15% MLM	83.33	84.67	77.00	73.67	72.67	61.67	75.84	82.22	70.13	90.00	91.00	87.00	80.71	82.64	73.95
100% NE+100%VB+15% MLM	83.00	86.67	77.67	77.67	74.33	65.00	81.21	83.56	75.84	89.00	88.67	84.00	82.72	83.31	75.63
100% NE+100%NN+15% MLM	82.67	85.33	75.00	75.33	72.67	62.00	80.54	84.23	74.48	90.33	90.00	86.33	82.22	83.06	74.45
100% NE+100% VB +100%NN+15% MLM	83.00	83.33	78.00	74.67	73.67	64.67	80.87	83.89	76.17	91.33	92.67	88.67	82.47	83.39	76.88
LEM$_{mono}$+LEM$_{para}$															
100% NE+15% TLM on 15%MLM	83.00	85.33	76.33	79.67	78.33	70.00	83.89	85.91	79.87	91.00	93.33	89.00	84.39	85.73	78.80
100% VB+15% TLM on 15%MLM	87.00	86.67	81.67	80.67	79.00	72.33	83.89	85.57	79.87	91.67	92.33	88.67	85.81	85.89	80.63
100% NN+15% TLM on 15%MLM	85.00	86.67	79.67	79.33	77.00	69.00	83.89	86.24	80.54	91.33	94.00	89.67	84.89	85.98	79.72
100% NE+100% VB+15% TLM on 15%MLM	85.67	76.67	69.33	79.67	76.67	69.33	83.22	84.23	77.85	92.00	92.00	89.33	85.14	82.39	76.46
100% NE+100% NN+15% TLM on 15%MLM	84.67	85.00	77.67	81.00	80.00	72.33	81.98	84.23	77.85	90.00	92.00	87.67	84.41	85.31	78.88
100% NE+100% VB+ 100% NN+ 15% TLM on 15%MLM	85.00	85.33	80.00	78.67	78.33	70.00	84.23	**88.59**	**80.87**	90.00	93.67	88.33	84.47	86.48	79.80
15% TLM on 100%NE+15%MLM	87.00	86.33	81.33	81.33	80.00	71.67	81.21	84.23	77.52	92.67	91.33	89.00	85.55	85.47	79.88
100%NE+15% TLM on 100%NE+15%MLM	87.67	87.00	81.67	82.00	**81.33**	**73.00**	81.88	84.23	77.18	91.33	92.67	88.33	85.72	86.31	80.05
100% VB+15% TLM on 100%NE+15%MLM	**88.33**	89.33	**83.67**	80.00	77.67	69.67	81.54	84.23	75.50	91.00	93.00	89.00	85.22	86.06	79.46
100% NN+15% TLM on 100%NE+15%MLM	86.33	87.67	80.33	81.33	79.67	70.67	80.54	84.56	76.51	92.00	91.33	88.00	85.05	85.81	78.88
100% NE+100% VB+15% TLM on 100%NE+15%MLM/	84.67	85.67	78.00	82.33	76.67	70.33	80.54	83.22	76.85	89.33	92.67	87.67	84.22	84.56	78.21
100% NE+100% NN+15% TLM on 100%NE+15%MLM/	84.67	85.67	78.00	82.33	76.67	70.33	80.54	83.22	76.85	89.33	92.67	87.67	84.22	84.56	78.21
100%NE+100%VB+100%NN+15%TLM on 100%NE+15%MLM/	85.00	84.33	78.33	78.00	76.67	67.33	79.53	83.89	75.84	92.33	92.00	90.00	83.72	84.22	77.88
15% TLM on (100%VB+15%MLM)	88.00	88.67	83.67	82.00	79.00	72.33	84.90	84.90	80.54	**93.33**	93.00	**91.00**	**87.06**	86.39	**81.88**
100% NE+ 15% TLM on (100%VB+15%MLM)	84.00	87.67	79.00	78.67	**81.33**	71.00	82.22	85.57	78.86	90.67	93.33	88.33	83.89	**86.98**	79.30
100% VB+ 15% TLM on (100%VB+15%MLM)	86.00	88.67	80.67	81.33	78.33	70.67	82.22	84.56	76.85	91.33	92.33	87.67	85.22	85.97	78.96
100% NN+15% TLM on (100%VB+15%MLM)	86.33	85.33	80.33	79.67	79.00	70.33	82.22	84.90	77.52	90.33	93.67	88.00	84.64	85.72	79.05
100% NE+ 100% VB+ 15% TLM on (100%VB+15%MLM)	85.67	88.00	80.33	80.33	76.00	69.00	81.54	83.58	77.18	90.67	93.00	88.00	84.55	85.14	78.63
100% NE+ 100% NN+ 15% TLM on (100%VB+15%MLM)	87.33	87.67	81.67	78.00	78.00	68.67	81.54	83.89	76.85	91.00	92.00	87.67	84.47	85.39	78.71
100% NE+ 100% NN+ 100%VB+ 15% TLM on (100%VB+15%MLM)	86.33	87.00	80.67	78.33	76.67	67.33	82.89	84.56	77.85	90.67	92.33	87.33	84.55	85.14	78.30
15% TLM on (100%NN+15%MLM)	84.67	88.33	81.00	81.00	77.33	69.67	83.22	85.91	78.86	91.67	92.33	89.67	85.14	85.98	79.80
100% NE+ 15% TLM on (100%NN+15%MLM)	85.33	86.67	79.00	78.33	76.33	67.33	81.88	84.56	76.85	91.00	91.33	87.67	84.14	84.72	77.71
100% VB+15% TLM on (100%NN+15%MLM)	84.67	87.67	80.67	78.67	76.00	67.33	79.19	84.29	75.50	90.00	92.67	88.00	83.13	85.16	77.88
100% NN+ 15% TLM on (100%NN+15%MLM)	85.00	87.00	79.33	77.33	76.33	66.00	80.87	83.56	74.83	89.67	92.67	87.00	83.22	84.89	76.79
100% NE+ 100% VB+ 15% TLM on (100%NN+15%MLM)	82.33	86.00	77.67	78.33	74.33	65.00	81.21	85.91	76.51	62.67	65.00	61.00	76.14	77.81	70.04
100% NE+ 100% NN+ 15% TLM on (100%NN+15%MLM)	86.00	87.00	80.00	78.00	76.67	66.67	79.53	84.23	74.16	88.67	92.33	86.67	83.05	85.06	76.87
100% NE+ 100% NN+ 100%VB+ 15% TLM on (100%NN+15%MLM)	86.33	**90.00**	82.00	76.00	78.33	66.33	79.87	85.91	76.16	90.33	92.00	87.67	83.13	86.56	78.04
15% TLM on (100%NE+100%VB+15%MLM)	85.33	89.33	80.00	80.00	75.33	67.33	85.34	84.29	78.86	90.00	91.67	87.00	85.17	85.16	78.30
100% NE+ 15% TLM on (100%NE+100%VB+15%MLM)	86.00	87.33	80.33	80.33	78.33	71.33	82.22	81.88	75.50	89.00	93.00	88.00	84.39	85.14	78.79
100% VB+15% TLM on (100%NE+100%VB+15%MLM)	84.67	87.67	79.33	78.00	75.33	67.00	83.89	84.56	77.85	88.33	92.00	86.67	83.72	84.89	77.71
100% NN+ 15% TLM on (100%NE+100%VB+15%MLM)	86.00	86.67	79.00	81.00	75.67	67.00	83.89	84.56	78.52	92.00	93.00	89.33	85.72	84.97	78.46
100% NE+100% VB+ 15% TLM on (100%NE+100%VB+15%MLM)	84.67	87.67	78.00	78.33	75.67	68.00	**90.87**	84.90	76.17	89.00	92.00	86.67	85.72	85.06	77.21
100% NE+ 100% NN+ 15% TLM on (100%NE+100%VB+15%MLM)	83.00	86.67	78.33	**84.67**	77.00	72.00	81.88	84.56	77.18	89.00	93.00	87.67	84.64	85.31	78.80
100%NE+100%VB+100%NN+15%TLM on (100%NE+100%VB+15%MLM)	87.67	86.33	80.00	79.00	78.00	69.33	82.89	84.29	78.52	90.00	93.67	88.67	84.89	85.57	79.13
15% TLM on (100%NE+100%NN+15%MLM)	86.00	89.33	80.00	80.00	76.33	69.67	85.34	85.24	78.86	90.00	92.67	88.00	85.33	85.89	79.13
100% NE+ 15% TLM on (100%NE+100%NN+15%MLM)	86.00	86.67	80.00	78.33	76.67	65.00	82.55	85.57	78.52	90.67	93.00	88.33	84.39	85.48	77.96
100% VB+15% TLM on (100%NE+100%NN+15%MLM)	84.67	87.67	79.33	80.67	78.67	70.00	81.54	85.23	78.86	89.33	93.00	87.00	84.05	86.14	78.80
100% NN+ 15% TLM on (100%NE+100%NN+15%MLM)	85.67	85.00	78.00	80.00	76.33	68.67	81.21	84.56	77.18	92.33	93.00	89.33	84.80	84.72	78.29
100% NE+100% VB+ 15% TLM on (100%NE+100%NN+15%MLM)	84.33	85.33	77.33	79.00	77.67	69.00	84.56	85.91	80.20	91.00	92.67	89.00	84.72	85.39	78.88
100% NE+100% NN+ 15% TLM on (100%NE+100%NN+15%MLM)	86.67	83.67	78.67	76.33	78.00	66.00	82.55	85.57	78.19	91.33	91.67	87.67	84.22	84.73	77.63
100%NE+100%VB+100%NN+15%TLM on (100%NE+100%NN+15%MLM)	82.33	86.00	76.67	77.00	79.00	68.67	81.21	82.55	75.84	91.00	91.33	88.00	82.89	84.72	77.29
15% TLM on (100%NE+100%VB+100%NN+15%MLM)	84.00	86.00	79.00	83.00	77.33	71.67	82.55	85.23	77.85	90.33	**94.33**	88.67	84.97	85.73	79.30
100% NE+ 15% TLM on (100%NE+100%VB+100%NN+15%MLM)	82.33	86.33	77.67	80.00	77.33	68.67	81.88	84.56	76.85	88.67	91.67	86.67	83.22	84.97	77.46
100% VB+15% TLM on (100%NE+100%VB+100%NN+15%MLM)	84.67	88.00	79.67	79.00	77.00	68.00	83.89	84.90	78.52	91.00	94.00	90.00	84.64	85.97	79.05
100% NN+ 15% TLM on (100%NE+100%VB+100%NN+15%MLM)	85.33	87.00	81.33	77.00	74.67	66.33	82.55	84.90	78.86	89.33	93.00	87.33	83.55	84.89	78.46
100% NE+100% VB+ 15% TLM on (100%NE+100%VB+100%NN+15%MLM)	82.67	85.67	78.33	76.67	75.00	66.00	83.22	85.91	77.52	88.00	92.67	85.67	82.64	84.81	76.88
100% NE+ 100% NN+ 15% TLM on (100%NE+100%VB+100%NN+15%MLM)	84.33	84.00	78.00	76.67	77.00	67.33	83.21	85.57	78.19	87.00	93.00	85.00	82.80	84.89	77.13
100%NE+100%VB+100%NN+15%TLM on (100%NE+100%VB+100%NN+15%MLM)	85.33	85.33	79.67	76.33	76.00	67.00	82.22	84.90	77.12	89.00	94.00	88.33	83.22	85.06	78.03

Table 11: LEM Ablation experiments and sentence alignment scores for En-Ta language pair considering linguistic entity masking. The best result for each criterion is shown in bold and highlighted in yellow.

Experiment	Army			Hiru			ITN			Newsfirst			Average		
	FW	BW	IN	FW	BW	IN	FW	BW	IN	FW	BW	IN	FW	BW	IN
Baselines															
XLM-R	83.44	81.46	78.15	90.67	91.00	87.33	91.33	90.00	87.00	**93.67**	95.33	**92.33**	89.78	89.45	86.20
15%MLM	86.75	88.08	81.46	88.00	89.33	84.00	93.33	92.67	89.33	90.33	94.00	89.00	89.60	91.02	85.95
15%TLM on 15%MLM	87.75	90.40	83.11	88.67	93.33	86.33	93.00	94.33	90.00	91.33	94.33	89.67	90.19	93.10	87.28
LEM$_{mono}$															
100% NE+15% MLM	86.42	92.05	83.78	89.33	92.00	87.67	94.00	94.33	90.67	91.33	94.00	88.67	90.27	93.10	87.69
100% VB+15% MLM	83.44	88.08	78.81	87.33	90.33	83.33	92.33	94.00	88.00	90.00	92.00	87.67	88.28	91.10	84.45
100% NN+15% MLM	85.10	87.75	80.13	88.00	91.67	85.33	92.00	91.67	88.00	90.67	93.33	87.67	88.94	91.10	85.28
100% NE+100%VB+15% MLM	84.43	90.73	82.12	88.67	91.00	85.33	94.00	92.33	88.33	91.00	94.33	88.00	89.53	92.10	85.95
100% NE+100%NN+15% MLM	85.43	88.08	79.47	88.33	89.67	85.00	95.00	94.67	91.33	92.33	93.33	89.67	90.27	91.44	86.37
100% NE+100% VB +100%NN+15% MLM	83.11	88.41	79.80	86.67	91.33	84.33	91.67	89.67	85.33	90.33	93.67	88.33	87.94	90.77	84.45
LEM$_{mono}$**+LEM**$_{para}$															
100% NE+15% TLM on 15%MLM	89.07	90.73	85.10	89.33	91.00	85.67	95.67	94.67	92.67	91.00	93.33	88.33	91.27	92.43	87.94
100% VB+15% TLM on 15%MLM	88.41	91.00	84.77	87.67	91.00	85.67	93.67	93.67	90.67	92.33	93.00	90.00	90.52	92.17	87.78
100% NN+15% TLM on 15%MLM	88.74	90.07	84.44	89.67	91.67	86.67	94.67	93.33	90.67	92.00	91.67	87.33	91.27	91.68	87.28
100% NE+100% VB+15% TLM on 15%MLM	86.75	90.73	83.11	89.67	90.33	86.33	92.67	92.67	88.67	92.33	**95.33**	90.67	90.36	92.27	87.19
100% NE+100% NN+15% TLM on 15%MLM	86.42	90.73	83.11	87.33	89.33	84.00	94.67	93.33	91.67	92.00	93.67	88.33	90.11	91.77	86.78
100% NE+100% VB+ 100% NN+ 15% TLM on 15%MLM	85.43	91.39	81.79	89.00	92.33	86.67	93.67	93.33	88.67	90.33	93.00	87.00	89.61	92.51	86.03
15% TLM on 100%NE+15%MLM	87.09	89.73	83.44	89.33	92.00	86.33	94.33	92.67	89.33	92.00	93.67	89.67	90.69	92.02	87.19
15% NE+15%TLM on 100%NE+15%MLM	88.33	**93.33**	87.33	88.33	93.33	87.33	93.33	94.00	89.00	92.00	93.67	89.67	90.50	**93.58**	88.33
100% VB+15% TLM on 100%NE+15%MLM	86.42	90.07	83.11	90.00	92.00	87.67	94.33	93.00	90.33	92.00	94.67	90.00	90.69	92.43	87.78
100% NN+15% TLM on 100%NE+15%MLM	86.09	91.72	83.78	89.67	92.67	87.67	95.33	95.00	91.67	92.67	93.33	89.67	90.94	93.18	88.19
100% NE+100% VB+15% TLM on 100%NE+15%MLM/	87.09	90.07	84.11	89.00	91.33	86.67	95.67	94.67	91.67	90.67	92.33	88.67	90.61	92.10	87.78
100% NE+100% NN+15% TLM on 100%NE+15%MLM/	86.09	90.73	83.11	90.00	**93.67**	**89.33**	94.33	94.00	90.33	91.00	**95.33**	89.33	90.36	93.43	88.03
100%NE+100%VB+100%NN+15%TLM on 100%NE+15%MLM/	86.09	93.05	84.11	89.67	92.00	88.00	95.67	95.00	**93.33**	91.00	94.00	89.33	90.61	93.51	88.69
15% TLM on (100%VB+15%MLM)	89.07	88.41	83.44	89.67	92.33	87.00	93.33	93.67	90.00	91.00	91.67	87.33	90.77	91.52	86.94
100% NE+ 15% TLM on (100%VB+15%MLM)	87.75	88.74	83.11	90.00	91.00	86.33	95.00	94.67	91.67	93.00	91.67	88.00	91.44	91.52	87.28
100% VB+ 15% TLM on (100%VB+15%MLM)	88.74	90.75	84.44	89.67	92.00	86.67	92.67	94.33	89.33	92.33	93.33	89.33	90.85	92.60	87.44
100% NN+15% TLM on (100%VB+15%MLM)	86.42	90.73	84.11	89.67	92.33	86.33	91.33	93.33	88.00	92.00	94.00	89.67	89.86	92.60	87.03
100% NE+ 100% VB+ 15% TLM on (100%VB+15%MLM)	85.76	88.08	81.13	90.33	92.33	88.00	93.00	91.33	88.67	92.33	92.00	88.33	90.36	90.94	86.53
100% NE+ 100% NN+ 15% TLM on (100%VB+15%MLM)	87.09	89.07	82.78	88.67	92.33	85.00	93.33	93.67	89.67	92.00	91.33	87.00	90.27	91.60	86.11
100% NE+ 100% NN+ 100%VB+ 15% TLM on (100%VB+15%MLM)	85.77	90.40	83.11	89.67	92.00	86.00	92.33	93.67	88.67	92.00	92.00	88.00	89.94	92.02	86.44
15% TLM on (100%NN+15%MLM)	88.41	91.39	85.76	88.33	92.67	85.67	95.67	**95.67**	91.67	91.00	93.67	89.33	90.85	93.35	88.11
100% NE+ 15% TLM on (100%NN+15%MLM)	**89.40**	92.72	87.42	90.13	93.33	88.67	96.67	93.00	90.67	92.33	93.67	89.67	**92.13**	93.18	89.10
100% VB+15% TLM on (100%NN+15%MLM)	87.75	90.07	83.11	87.67	92.00	85.00	93.33	93.33	89.00	89.67	92.33	87.67	89.60	91.93	86.19
100% NN+ 15% TLM on (100%NN+15%MLM)	85.43	90.73	82.12	88.67	93.00	86.67	95.33	93.67	91.00	93.00	92.67	89.33	90.61	92.52	87.28
100% NE+ 100% VB+ 15% TLM on (100%NN+15%MLM)	86.09	90.40	82.78	**91.33**	92.00	88.33	94.67	94.67	91.33	91.33	93.00	88.33	90.86	92.52	87.69
100% NE+ 100% NN+ 15% TLM on (100%NN+15%MLM)	87.75	89.40	83.11	88.33	92.00	85.67	95.00	93.67	90.67	92.00	93.00	89.33	90.77	92.02	87.19
100% NE+ 100% NN+ 100%VB+ 15% TLM on (100%NN+15%MLM)	85.43	92.05	83.78	89.33	93.00	87.00	94.33	93.33	90.00	90.67	92.67	88.00	89.94	92.76	87.19
15% TLM on (100%NE+100%VB+15%MLM)	86.09	91.06	83.44	90.33	90.67	87.33	96.33	94.33	92.33	92.33	94.33	90.00	91.27	92.60	88.28
100% NE+ 15% TLM on (100%NE+100%VB+15%MLM)	86.75	90.07	83.44	89.33	91.33	86.67	93.67	94.67	91.33	92.00	93.33	89.67	90.44	92.35	87.78
100% VB+15% TLM on (100%NE+100%VB+15%MLM)	84.44	91.39	81.46	89.33	91.67	85.00	95.33	93.33	91.00	92.00	93.33	89.33	90.28	92.43	86.70
100% NN+ 15% TLM on (100%NE+100%VB+15%MLM)	85.76	92.05	85.00	90.67	93.00	88.33	95.67	95.33	92.67	89.33	92.67	86.33	90.36	93.26	88.08
100% NE+100% VB+ 15% TLM on (100%NE+100%VB+15%MLM)	87.09	89.73	83.11	90.67	92.33	88.33	94.67	92.67	89.33	92.00	93.00	90.00	91.10	91.93	87.69
100% NE+ 100% NN+ 15% TLM on (100%NE+100%VB+15%MLM)	85.76	92.05	85.00	90.67	91.67	87.67	95.67	94.00	91.33	90.67	93.67	88.33	90.69	92.85	88.08
100%NE+100%VB+100%NN+15%TLM on (100%NE+100%VB+15%MLM)	86.75	92.05	84.77	88.33	90.00	86.67	94.33	94.67	90.67	89.67	93.67	87.67	89.77	92.60	87.44
15% TLM on (100%NE+100%NN+15%MLM)	86.75	91.72	83.11	90.67	92.33	88.67	95.67	93.67	91.33	92.00	94.33	90.00	91.27	93.01	88.28
100% NE+ 15% TLM on (100%NE+100%NN+15%MLM)	86.75	87.47	81.79	90.33	93.00	88.00	95.33	93.33	90.33	**93.67**	94.00	90.33	91.52	91.95	87.61
100% VB+15% TLM on (100%NE+100%NN+15%MLM)	87.75	90.40	83.44	89.33	92.33	86.67	95.00	94.00	91.33	91.67	94.67	89.67	90.94	92.85	87.78
100% NN+ 15% TLM on (100%NE+100%NN+15%MLM)	87.75	90.40	84.77	87.67	89.67	84.00	95.33	95.33	92.00	90.00	93.67	88.67	90.19	92.27	87.36
100% NE+100% VB+ 15% TLM on (100%NE+100%NN+15%MLM)	85.76	87.75	81.13	90.33	90.67	86.00	94.33	92.00	89.00	92.33	93.67	90.00	90.69	91.02	86.53
100% NE+100% NN+ 15% TLM on (100%NE+100%NN+15%MLM)	87.75	90.07	84.44	90.33	92.00	86.67	96.00	94.00	91.67	93.00	93.67	89.00	91.77	92.43	87.94
100%NE+100%VB+100%NN+15%TLM on (100%NE+100%NN+15%MLM)	85.76	88.74	**91.46**	89.67	92.00	86.00	94.33	93.67	91.00	93.00	94.00	89.67	90.69	92.10	**89.53**
15% TLM on (100%NE+100%VB+100%NN+15%MLM)	86.09	89.40	82.45	89.33	92.00	87.00	94.33	91.67	88.33	91.33	92.00	86.33	90.27	91.27	86.03
100% NE+ 15% TLM on (100%NE+100%VB+100%NN+15%MLM)	88.76	89.40	84.44	89.67	91.33	87.33	95.00	94.00	90.67	90.33	92.00	87.00	90.94	91.68	87.36
100% VB+15% TLM on (100%NE+100%VB+100%NN+15%MLM)	85.76	89.40	82.12	90.33	93.33	88.67	94.67	94.00	90.67	93.00	93.67	90.00	90.94	92.60	87.86
100% NN+ 15% TLM on (100%NE+100%VB+100%NN+15%MLM)	85.43	90.73	82.78	89.67	91.67	87.33	95.33	92.33	90.33	90.00	93.33	86.67	90.11	92.02	86.78
100% NE+100% VB+ 15% TLM on (100%NE+100%VB+100%NN+15%MLM)	86.42	91.00	82.78	88.67	91.67	86.33	94.00	92.67	89.33	93.00	93.67	90.33	90.52	92.25	87.19
100% NE+ 100% NN+ 15% TLM on (100%NE+100%VB+100%NN+15%MLM)	86.75	90.73	84.11	90.67	93.33	88.33	**97.33**	92.67	92.33	91.33	92.00	86.67	91.52	92.18	87.86
100%NE+100%VB+100%NN+15%TLM on (100%NE+100%VB+100%NN+15%MLM)	87.42	89.40	82.78	89.33	92.33	86.33	94.67	95.00	91.33	91.33	94.00	88.67	90.69	92.68	87.28

Table 12: LEM Ablation experiments and sentence alignment scores for Si-Ta language pair considering linguistic entity masking. The best result for each criterion is shown in bold and highlighted in yellow.