Modular Sentence Encoders: Separating Language Specialization from Cross-Lingual Alignment

Anonymous ACL submission

Abstract

Multilingual sentence encoders (MSEs) are commonly obtained by training multilingual language models to map sentences from different languages into a shared semantic space. As such, they are subject to *curse of multilinguality*, a loss of monolingual representational accuracy due to parameter sharing. Another limitation of MSEs is the trade-off between different monolingual and cross-lingual performance: training for cross-lingual alignment of sentence embeddings distorts the optimal monolingual structure of semantic spaces of individual languages, harming the utility of sentence embeddings in monolingual tasks; cross-lingual tasks, such as cross-lingual semantic similarity and zero-shot transfer for sentence classification, thus may require different kind of cross-lingual alignment training. In this work, we address both issues by means of modular training of sentence encoders. We first train language-specific monolingual modules to mitigate negative interference between languages (i.e., the curse). We then align all non-English sentence embeddings to the English by training cross-lingual alignment adapters, preventing interference with monolingual specialization from the first step. We train and merge two types of cross-lingual adapters to resolve the conflicting requirements of different cross-lingual tasks. Monolingual and cross-lingual results on semantic text similarity and relatedness, bitext mining and sentence classification tasks show that our modular solution achieves better and more balanced performance across all the tasks compared to full-parameter training of monolithic multilingual sentence encoders, especially benefiting low-resource languages.[1]

1 Introduction

Multilingual Sentence Encoders (MSEs; Artetxe and Schwenk, 2019b; Yang et al., 2020; Reimers and Gurevych, 2020; Feng et al., 2022; Duquenne et al., 2023) embed sentences from different languages into a shared semantic vector space, making them essential tools for multilingual and cross-lingual semantic retrieval (e.g., bitext mining; Schwenk et al., 2021), clustering (e.g., for extractive summarization; Bouscarrat et al., 2019), and filtering (e.g., in content-based recommendation; Hassan et al., 2019), as well as for cross-lingual transfer in supervised text classification (Artetxe and Schwenk, 2019b; Licht, 2023). In this work, we aim to address two limitations in the MSEs through modular training: the curse of multilinguality and the trade-off in performance between different monolingual and cross-lingual tasks.

Like general-purpose multilingual encoder language models (mELMs, e.g., mBERT; Devlin et al., 2019, XLM-R; Conneau et al., 2020), multilingual models specialized for sentence encoding[2] are also subject to the *curse of multilinguality* (Conneau et al., 2020), a loss of representational precision for each individual language due to sharing of model parameters between many languages, resulting in negative interference (Wang et al., 2020). Training language-specific modules like embedding layers and language adapters (Pfeiffer et al., 2021, 2022) or full models (Blevins et al., 2024) has been proven effective against this issue for general-purpose models, but rarely applied for MSEs, whose sentence embeddings from different monolingual modules need to be semantically aligned to each other. To the best of our knowledge, the only work that targets CoM for MSEs is LASER3 (Heffernan et al., 2022): they train a set of monolingual sentence encoders from scratch through the distillation from a fixed teacher MSE, which is already affected by the CoM.

Existing MSE work mostly focuses on cross-lingual training and evaluation, paying less atten-

[1] Our code is available in Supplementary Material.

[2] In fact, many MSEs are derived from mELMs (Reimers and Gurevych, 2020; Feng et al., 2022, *inter alia*) by doing sentence-level training on top of them.

tion to the monolingual (i.e., within-language) performance, which can be negatively affected by the cross-lingual alignment (Roy et al., 2020). Earlier work on inducing cross-lingual word embeddings (Søgaard et al., 2018; Patra et al., 2019; Glavaš and Vulić, 2020) hints at an explanation for this trade-off: forcing cross-lingual alignment between non-isomorphic monolingual spaces distorts those spaces and thus degrades their monolingual semantic quality. What is more, there also seems to be a tradeoff between different cross-lingual tasks and different cross-lingual training approaches yield optimal performance for different tasks. Concretely, MSEs trained on *parallel* data to produce highly similar embeddings for exact translation pairs are effective in bitext mining (Artetxe and Schwenk, 2019b; Feng et al., 2022; Heffernan et al., 2022); however, they perform worse on cross-lingual semantic similarity, failing to produce high similarity for sentences with *similar* but non-equivalent meaning (Reimers and Gurevych, 2020). Conversely, MSEs trained on cross-lingual *paraphrase* data (Yang et al., 2020; Wang et al., 2022), i.e. pairs of semantically similar but non-equivalent sentences, yield better semantic similarity performance but are not effective in bitext mining. Paraphrase-trained models also seem to offer weaker performance in zero-shot cross-lingual transfer for sentence classification tasks (Roy et al., 2020), which also seems to benefit more from parallel alignment.

Contributions. In this work, we propose to alleviate all of the above shortcomings by means of *modularity*, that is, parameter separation. As illustrated in Figure 1, we first mitigate the curse of multilinguality by specializing an MSE for each target language, by training language-specific embedding layers and language adapters via masked language modeling (MLM-ing). To obtain high-quality monolingual sentence embeddings, we then train a monolingual sentence encoding adapter (SE adapter) for each language on top of the language adapter, resorting to sentence-level contrastive learning on synthetic monolingual paraphrase data, machine-translated from English. In the next step, we carry out cross-lingual alignment training also in a modular fashion, without jeopardizing the monolingual sentence representation quality. Further, to meet the requirements of different cross-lingual tasks, we train two kinds of cross-lingual alignment adapter (CLA adapter) for each language—one is trained on cross-lingual *paraphrase* data, and the other on *parallel* data—and merge them post-hoc into a single CLA adapter by means of weight averaging: this offers the flexibility of arbitrary weighting between the two types of training approaches for different cross-lingual sentence-level tasks. At inference time, we activate the language-specific modules (embeddings, language adapter, SE adapter, CLA adapter) of the respective language of the input sentence.

Our experiments—encompassing four tasks and 23 linguistically diverse languages and two state-of-the-art MSE models—render our modular approach effective in overcoming the performance trade-offs between both (1) monolingual and cross-lingual tasks as well as (2) different sentence-level tasks types (semantic textual similarity and relatedness on the one side vs. bitext mining and sentence classification on the other), with substantial performance gains over full-parameter training of a single monolithic MSE. Our approach particularly benefits low-resource languages, most affected by the curse of multilinguality. Since both contrastive learning steps in our approach—for monolingual specialization and for cross-lingual alignment—are carried out on machine-translated data, our work also validates the viability of MT for scaling up MSE training data.

2 Related Work

Multilingual Sentence Embeddings. Multilingual sentence encoders should produce similar sentence embeddings for sentences with similar meaning, regardless whether they come from the same or different languages. Cross-lingual alignment is thus at the core of MSE training, typically achieved by training on parallel data with a translation objective (Artetxe and Schwenk, 2019b; Duquenne et al., 2023; Gao et al., 2023), or contrastive loss (Feng et al., 2022; Zhao et al., 2024). As a standard practice to acquire high-quality English sentence embedding(Reimers and Gurevych, 2019; Gao et al., 2021), contrastive learning with *paraphrase* pairs[3] has also been applied to train MSEs. This can be done through teacher-student distillation with an English teacher model trained on English paraphrases (Reimers and Gurevych, 2020; Ham and Kim, 2021), or directly with cross-lingual paraphrases (Wang et al., 2022, 2024). Another line of work removes language-specific information to get

[3]We use the word "paraphrase" in a broad sense, to include also, e.g., entailment pairs or question-answer pairs.

language-agnostic meaning representation (Yang et al., 2021; Tiyajamorn et al., 2021; Kuroda et al., 2022). To the best of our knowledge, our work is the first attempt to address multiple conflicting factors in MSE training, aiming to yield optimal performance trade-off across a variety of tasks.

Lifting the Curse of Multilinguality. Large body of work focuses on post-hoc parameter-efficient adaptation of multilingual models for individual languages (Pfeiffer et al., 2020, 2021; Parović et al., 2022, *inter alia*) through continued pretraining on the target language corpora. Expanding or replacing multilingual vocabulary with target language tokens and smart initialization of their embeddings (Chau and Smith, 2021; Pfeiffer et al., 2021; Minixhofer et al., 2022; Dobler and de Melo, 2023) has been shown to improve sample efficiency of post-hoc language adaptation of multilingual models.

While language-specific modular training is a common approach for post-hoc adaptation of vanilla language models, it is seldom applied on MSEs, as MSE training additionally requires specialization for sentence encoding and cross-lingual alignment. Existing language-specific sentence encoders still rely on monolithic full-parameter training of the whole model: they are either trained only for a certain language (Mohr et al., 2024), or distilled from a massively multilingual teacher model which is already affected by the curse of multilingual and never really trained to model fine-grained semantic similarity (Heffernan et al., 2022). Some existing MSE efforts (Mao et al., 2021; Kuroda et al., 2022; Liu et al., 2023; Yano et al., 2024) do leverage lightweight modules for cross-lingual training, but these modules are still (massively) multilingual, i.e., do not alleviate the curse of multilinguality.

3 Modular Sentence Encoder

Our main objective is to obtain multilingual sentence embeddings that excel across the board, despite the conflicts between different tasks and scenarios: (i) in both monolingual and cross-lingual tasks, despite cross-lingual semantic alignment possibly being at odds with monolingual semantic specialization; and (ii) in different types of cross-lingual tasks, despite the fact that they require different types of cross-lingual alignment training (Roy et al., 2020). To mitigate these inherent trade-offs, we propose a modular approach, i.e., to isolate parameters for each requirement, as illustrated in Figure 1. We train a set of language-specific modules to (1) specialize the MSE for each individual language, and (2) to align the monolingually adapted MSEs for cross-lingual tasks. To create training data for every language, we machine-translate a mixture of English paraphrase datasets.

Monolingual Specialization. We specialize MSEs like LaBSE (Feng et al., 2022) and mE5 (Wang et al., 2024) for each language by training language-specific (i) embedding layers and (ii) adapters, using monolingual data.

Language Adaptation (LA). For each language, we train a new, language-specific tokenizer and initialize its new embedding matrix following the FOCUS approach (Dobler and de Melo, 2023). In a nutshell, FOCUS copies the embeddings for tokens that already exist in the vocabulary of the original MSE; for new tokens, it interpolates between embeddings of similar tokens from the original vocabulary. Compared to random initialization, FOCUS keeps a substantial amount of information from the pre-trained embeddings of the multilingual model in the new embeddings, making them "compatible" with the model body, avoiding the need to train them from scratch for each language: this leads to more sample efficient training for the embedding layers.[4] For each target language, we then do standard (continued) MLM-ing on the monolingual corpora of the language. To this end, we resort to modular, parameter-efficient fine-tuning (PEFT): besides the parameters of the new embedding matrix, we train only the low-rank adaptation matrices (LoRA; Hu et al., 2022) in encoder's layers. PEFT has been widely adopted for post-hoc language specialization of vanilla mELMs (Pfeiffer et al., 2020, 2021; Parović et al., 2022).

(Re-training for) Sentence Encoding (SE). As a token-level objective, (continued) MLM-ing is detrimental to the original sentence embedding abilities of a pre-trained MSE: we thus need to re-specialize each language-specific encoder for (monolingual) sentence encoding: for this, we use a standard contrastive learning objective, Multiple Negative Ranking Loss (MNRL; Henderson et al., 2017) and train on the (noisy) monolingual paraphrase data, machine-translated from English. This step is also done in a modular way by stacking another set of monolingual adapters (again LoRA), the *SE adapter*, on top of the LA. In this training step, only the parameters of the SE

[4]We refer the reader to the original paper for more details.

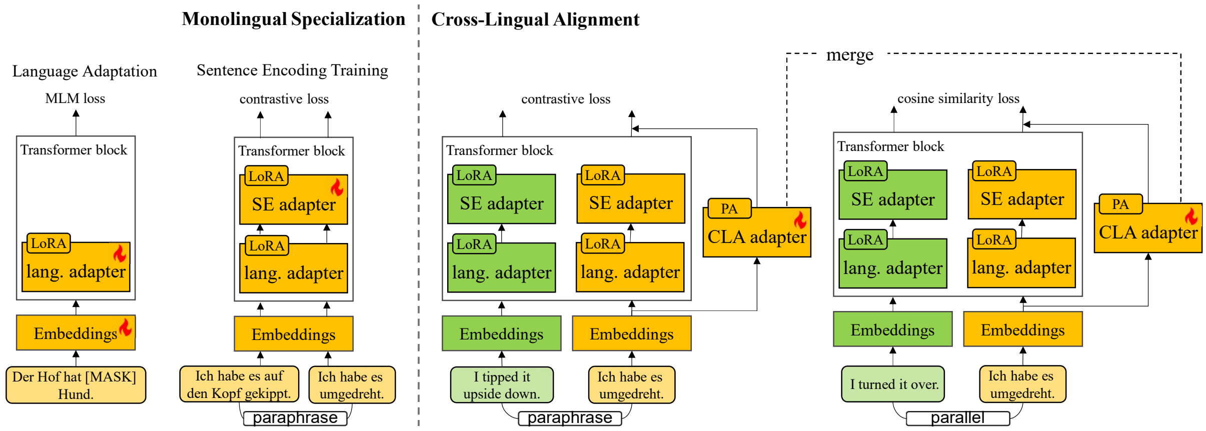

Figure 1: Illustration of how we apply our modular training to a pre-trained multilingual sentence encoder. In each step, only the module marked with the fire symbol is trained. In the monolingual specialization step, we train a language-specific embedding layer, a language adapter and a monolingual sentence encoding (SE) adapter for each language. In the cross-lingual alignment (CLA) step, the monolingual representation is aligned to the English representation via paraphrase and parallel data, respectively. The two CLA adapters are merged through weight averaging to form the final CLA adapter for each language. PA stands for parallel adapter.

adapter are trained, in order to obtain the monolingual sentence encoding ability; the encoder body, language-specific embeddings layer and the previously trained LA are all kept frozen.

Cross-Lingual Alignment (CLA). The mutually independent language adaptation for individual languages warrants a cross-lingual sentence-level alignment step, so that the sentence embeddings can also be used in cross-lingual applications. To prevent negative interference between cross-lingual alignment and previously imparted monolingual SE abilities, we train a cross-lingual alignment (CLA) module as a *parallel adapter* (He et al., 2022) for each non-English language. The cross-lingual alignment training then updates only language-specific CLA adapters: the monolingual modules of the corresponding input language are activated in the forward pass, but not updated.

Since our machine-translated monolingual paraphrase datasets are parallel across all languages, we can create two sorts of cross-lingual training data: *paraphrase* pairs (i.e. sentence in language A and its paraphrase in language B) and *parallel* pairs (i.e. sentence in language A and its direct translation in language B). We mitigate the inherent interference between bitext mining and semantic similarity (see §1), by training two separate CLA adapters for each language, one on *parallel* and one on *paraphrase* data. Specifically for each language other than English, we train (1) the *paraphrase* CLA adapter with the MNRL loss—just like the SE adapter in monolingual SE specialization—only now with bilingual (English-target language) paraphrase pairs; in contrast, we train (2) the *parallel* CLA adapter with the cosine similarity loss (following (Heffernan et al., 2022)) on bilingual (English-target) translation pairs. One can then either use the more suitable of the two CLA adapters in a downstream tasks or post-hoc interpolate between the two skills (i.e., ability to model fine-grained cross-lingual semantic similarity and the ability to match representations of exact translations) to best match the what is needed for a concrete downstream task.

We favor bilingual alignment with the English encoder over multilingual alignments[5], because English embeddings are the most reliable: not only is the initial multilingual encoder most "fluent" in English, but we also trained English embeddings on gold paraphrase data, whereas all other SEs are trained with noisy translations. Because of this, we omit to train the CLA adapter(s) for English: with English embedding space being of best semantic quality, we want embeddings from other languages to adapt (through their CLA adapters) to the English space, and not vice versa. Using English as a pivot has already been proven effective in aligning non-English languages to each other (Reimers and Gurevych, 2020; Heffernan et al., 2022).

Inference. After training, we have several modules for each (target) language: embedding layer, language adapter, SE adapter and CLA adapter(s).

[5]Given the multi-parallel nature of the paraphrase data we obtained with MT, direct alignment between all non-English language pairs is possible.

When encoding the input text, the corresponding modules for the input language should be activated. Thus, the language of the input text should be known. Otherwise, one can easily apply any SotA language identification models (Kargaran et al., 2023) to detect the input language first.

4 Experimental Setup

Models. We start from two popular MSEs as base models for our modular specialization: **LaBSE** and multilingual E5 (**mE5**-base). LaBSE has been trained on billions of parallel sentence pairs (Feng et al., 2022). Starting from XLM-R (Base) (Conneau et al., 2020), mE5-base has first been trained on around 1 billion of (noisy) weak-supervision pairs, then on around 1.6 million high-quality sentence pairs (Wang et al., 2024). The goal of our work is *not* to outperform *other* MSEs or achieve SotA performance; instead, we aim to show that our proposed modular specialization offers clear benefits over monolithic single-model training.

Monolithic Baselines. Our primary baseline is the single monolithic MSE model for which all parameters are updated in each training step, akin to mSimCSE (Wang et al., 2022). While mSimCSE originally trains only on (English or cross-lingual) NLI data, we extend this to make the comparison with our modular variants as fair as possible: we use all of the MT-obtained multilingual paraphrase instances as in our modular training. We have the following monolithic-model variants: (i) $\mathbf{Full_{en}}$, trained only on (clean) English paraphrase data; (ii) $\mathbf{Full_{m}}$, trained only on monolingual data of all languages (each batch is monolingual, language randomly sampled for each batch); (iii) $\mathbf{Full_{c}}$, trained only on cross-lingual paraphrase pairs (the language for each sentence in a paraphrase pair is randomly selected); and (iv) $\mathbf{Full_{mc}}$, trained sequentially, first on monolingual (m) and then on cross-lingual (c) paraphrases.

Modular Variants. We evaluate the following variants: (i) $\mathbf{Mod_{en}}$, as a baseline: a monolingual SE adapter is trained only on English paraphrase data and used for all other languages; i.e., we transfer the sentence encoding ability from English; (ii) $\mathbf{Mod_{m}}$: only monolingual specialization, i.e. a monolingual SE adapter is trained with paraphrase dataset for every language; (iii) $\mathbf{Mod_{mc\text{-}pp}}$ adds a CLA adapter trained on cross-lingual *paraphrase* data to $\mathrm{Mod_m}$; (iv) $\mathbf{Mod_{mc\text{-}pl}}$ adds a CLA adapter trained on cross-lingual *parallel* data to $\mathrm{Mod_m}$; (v) $\mathbf{Mod_{mc}}$ merges the two CLA adapters in $\mathrm{Mod_{mc\text{-}pp}}$ and $\mathrm{Mod_{mc\text{-}pl}}$ (with equal contribution) into one single CLA adapter. We do the modular training on LaBSE for 23 languages present in evaluation datasets. Due to the intensive LA step and limited resources, for mE5 we train the modules for a subset of 10 languages.[6]

Training Data. To get multilingual paraphrase data, we translate, with NLLB 3.3B as our MT model (NLLB Team et al., 2022), five English paraphrase datasets—MNLI (Williams et al., 2018), SentenceCompression (Filippova and Altun, 2013), SimpleWiki (Coster and Kauchak, 2011), Altlex (Hidey and McKeown, 2016) and QuoraDuplicateQuestions, containing combined around 600K sentence pairs (see Appendix C.1 for details on the datasets)—into all 22 languages found in our downstream evaluation datasets. This results in a multi-parallel paraphrase dataset spanning 23 languages, from which we create instances for monolingual and cross-lingual training.

We train language-specific tokenizers and carry out monolingual language adaptation on monolingual corpora combined from language-specific portions of CC100 (Conneau et al., 2020) and MADLAD-400 (Kudugunta et al., 2023).

Evaluation Data. We evaluate the obtained sentence encoders on four tasks: STS, STR, bitext mining, and sentence classification. For the first three tasks, we do evaluation in the "zero-shot" setup, i.e., without any task-specific supervised training. We only evaluate on high-quality datasets, compiled either manually from scratch or by human post-editing of machine translations.[7]

Semantic Textual Similarity (STS). The models need to produce a score indicating semantic similarity for a pair of sentences. We simply predict the cosine similarity between the embeddings of the sentences. Performance is reported as Spearman correlation ($\times$100) against human scores. We collect existing multilingual STS datasets and use parallel monolingual STS data to create high-qualiry cross-lingual evaluation pairs. For example, the STS datasets for Czech, German and French (Hercig and Kral, 2021) and the datasets for Dutch, Italian and Spanish (Reimers and Gurevych, 2020) are parallel to each other, as they are translated from the same **STS17** (Cer et al., 2017) English data.

[6]We provide the full list of languages in Appendix A and implementation and training details in Appendix B.

[7]See Appendix C.2 for more details on evaluation datasets.

The same applies for the STS datasets for Turkic languages in Kardeş-NLU (Senel et al., 2024) and the Korean STS dataset from Ham et al. (2020), all translated from the English STS-Benchmark (**STSB**; Cer et al., 2017). We can thus leverage this effectively multi-parallel STS data for cross-lingual evaluation on many more language pairs, including pairs never evaluated in prior work, e.g. Czech-Italian or Korean-Uzbek.

Semantic Textual Relatedness. Semantic relatedness is a broader concept than similarity, that also considers aspects like topic or view similarity (Ousidhoum et al., 2024). We use the same metric as in the STS task. Similar to STS, we aggregate the multi-parallel monolingual data and create cross-lingual pairs between Polish (Dadas et al., 2020), Dutch (Wijnholds and Moortgat, 2021), and Spanish (Araujo et al., 2022), all translated from the English **SICK** dataset (Marelli et al., 2014). **STR24** (Ousidhoum et al., 2024) contains monolingual STR data for low-resource African and Asian languages; but it is not multi-parallel, and as such only lends itself to monolingual evaluation.

Bitext Mining. The model should mine parallel sentences (translation pairs) from two lists of monolingual sentences based on the cosine similarity of bilingual sentence pairs. Following Heffernan et al. (2022), we use the `xsim` score (error rate of wrongly aligned sentences; Artetxe and Schwenk, 2019a) to evaluate our models on two bitext mining datasets: **FLORES** (Goyal et al., 2022) and **Tatoeba** (Artetxe and Schwenk, 2019b). Since FLORES is multi-parallel, we test on all possible language pairs between our target languages. Tatoeba only contains English-X data: we average the results from both mining directions (English→X and X→English) for all languages X.

Topic Classification. We resort to **SIB**-200 (Adelani et al., 2024) to obtain data for topical sentence classification for our target 23 languages. In monolingual evaluation, we train a simple Logistic Regression (Cox, 1958) classifier on top of a frozen sentence encoder for each target language. In (zero-shot) cross-lingual transfer setup, we train the classifier on English data.

Alignment metrics. In standard task formulations, cross-lingual STS and bitext mining are *bilingual*, i.e., a sentence in one language is compared only against sentences in one (and same) other language. Such an evaluation setup fails to capture the **language bias** of an MSE (Roy et al., 2020): in a multilingual candidate pool, the model might prefer certain language (pair) over others, e.g., map sentences from the same language closer in the embedding space even if they are semantically dissimilar. Following (Reimers and Gurevych, 2020), we quantify language bias as the performance drop when switching from bilingual to multilingual evaluation, in which we calculate the Spearman correlation on the concatenation of all bilingual datasets. To this end, we use the multi-parallel STSB and SICK datasets; we report the difference between the average performance on bilingual tasks and the performance on the single multilingual task. Another indicator of semantic quality of multilingual representation spaces is the similarity of monolingual semantic structures, i.e., the degree of their isomorphism. It can be quantified by Relational Similarity, (**RSIM**; Vulić et al., 2020) on a bilingual parallel corpus: we calculate the corresponding sets of cosine similarity scores for all monolingual sentence pairs, in each of the two languages and report RSIM as Pearson correlation between the two sets of corresponding monolingual cosines. We measure RSIM on FLORES, averaging the results across all language pairs.

5 Results and Discussion

We report the results for our LaBSE-based models in Table 1 and for mE5-based models in Table 2.

5.1 Full Model Results

Further training on monolingual paraphrase data (Full_{en} and Full_{m}) can already largely improve the original models' (first row in each table) performance on all tasks, except for bitext mining. The off-the-shelf LaBSE model is a strong baseline for bitext mining, as it has been pre-trained on a massive amount of parallel data, which perfectly aligns with the goal of bitext mining. This confirms previous finding that training on paraphrase data can disturb bitext mining ability (Reimers and Gurevych, 2020). Full_{m} trained on monolingual data of all target languages outperforms Full_{en} (i.e., the $\text{mSimCSE}_{\text{en}}$ baseline) slightly on LaBSE and significantly on mE5, demonstrating the limitation of cross-lingual transfer of sentence-embedding specialization from English, especially if the base model has not been subjected to massive cross-lingual pre-training on parallel data like LaBSE.

Full_{m} outperforms Full_{c}on monolingual STS/STR tasks, whereas the opposite is true in

	Monolingual tasks					Cross-lingual tasks						Alignment metrics		
	STS↑		**STR↑**		**CLS↑**	**STS↑**		**STR↑**	**CLS↑**	**Bitext Mining↓**		**Language Bias↓**		**RSIM↑**
Dataset	sts17	stsb	sick	str24	sib	sts17	stsb	sick	sib	flores	tatoeba	stsb	sick	flores
LaBSE	76.7	71.9	68.0	69.2	82.7	74.5	64.4	63.8	83.6	0.14	3.87	1.02	2.32	0.64
Full$_{en}$	82.7	**80.9**	**76.5**	75.4	84.1	78.8	71.5	70.4	83.5	0.49	4.72	0.87	1.48	0.70
Full$_{m}$	**82.9**	80.4	76.4	**75.9**	84.8	**79.4**	71.5	70.9	83.9	0.29	4.43	0.88	1.27	0.74
Full$_{c}$	81.0	79.1	75.1	75.3	85.1	77.8	72.1	71.5	85.3	**0.20**	**4.00**	**0.53**	0.70	**0.77**
Full$_{mc}$	80.0	79.2	75.1	75.4	**86.0**	76.7	**72.7**	**71.7**	**86.3**	0.21	4.17	**0.53**	**0.64**	**0.77**
Mod$_{en}$	82.6	82.1	76.3	**78.7**	84.9	80.1	74.8	71.5	83.6	0.16	3.68	0.90	1.24	0.73
Mod$_{m}$	**83.1**	82.1	76.5	78.4	85.5	80.6	75.3	71.9	85.0	**0.15**	3.63	1.05	1.16	0.75
Mod$_{mc\text{-}pp}$	82.9	81.8	**76.7**	77.5	86.0	**80.7**	76.0	**72.8**	85.0	0.16	**3.49**	0.71	0.92	0.76
Mod$_{mc\text{-}pl}$	81.4	81.6	76.0	77.2	85.8	79.1	76.1	72.4	**86.2**	**0.15**	3.64	**0.56**	**0.67**	**0.82**
Mod$_{mc}$	82.7	**82.2**	76.6	78.4	**86.1**	80.5	**76.4**	72.6	85.3	**0.15**	3.53	0.59	0.81	0.78
Ablations														
Mod$_{m}$−LA	81.3	78.1	74.3	75.9	84.0	79.0	72.0	71.0	84.7	0.13	3.84	0.85	1.10	0.75
Mod$_{c}$	82.7	81.6	76.0	79.0	85.7	80.7	75.6	72.0	85.1	0.14	3.54	1.04	1.61	0.75

Table 1: Results of the LaBSE-based models for 23 languages. Reported results are averages over all languages in each evaluation dataset. The best result within the Full group and the Mod group on each dataset is denoted in **bold**. The second-best result in the Mod group is underlined.

cross-lingual tasks: this confirms the inherent trade-off between monolingual and cross-lingual abilities of MSEs. The inability of monolingual training, even using multi-parallel data, to induce strong cross-lingual semantic structures is confirmed by the higher language bias and lower RSIM scores of Full$_{m}$. The trade-off between monolingual and cross-lingual performance is more pronounced in mE5 results. The sequential combination of both monolingual and cross-lingual training (Full$_{mc}$) is unable to resolve the conflict and yields results similar to Full$_{c}$: in a monolithic MSE model, the subsequent cross-lingual alignment seems to distort the semantic quality of monolingual subspaces. One notable exception is *monolingual* text classification where Full$_{mc}$performs the best. We speculate that is because topic classification relies on lexical cues rather than fine-grained sentence meaning: cross-lingual training probably improves lexical alignments and the fine-grained distortions it brings to monolingual semantics play no role in this semantically coarse task.

5.2 Modular Model Results

Monolingual Training. We first compare the baseline Mod$_{en}$, with an SE adapter trained only on English data against Mod$_{m}$ with a language-specific SE adapter. As is the case for monolithic models, Mod$_{m}$ with language-specific SE adapters trained with noisy machine-translated data, outperforms the transfer from English-only SE training (Mod$_{en}$), drammatically reducing the language bias for mE5. Looking at performance on monolingual tasks, our Mod$_{m}$ with monolingual (LA and SE) specialization successfully mitigates the curse of multilinguality, which seems to be present in its monolothic counterpart Full$_{m}$: the gains are particularly prominent on STSB (+1.7 on LaBSE, +2.5 on mE5) and STR24 (+2.5 on LaBSE), datasets that encompass most low-resource languages.

The importance of modularity becomes most apparent on *cross-lingual* STS, where our Mod$_{m}$, not exposed to any explicit cross-lingual alignment outperforms the explicitly cross-lingually trained monolithic variants (Full$_{c}$ and Full$_{mc}$). This shows that monolingual training on multi-parallel data leads to semantic alignment, emphasizing the potential of MT for synthesizing MSE training data. Adding cross-lingual alignment in a modular fashion (Mod$_{mc}$ variants) brings further gains (compared to Mod$_{m}$) in sentence classification transfer (CLS) and reduces the language bias. Crucially, Mod$_{mc}$ variants slightly outperform Mod$_{m}$on *monolingual* tasks, showing that modularity mitigates the conflict between monolingual and cross-lingual performance that MSEs suffer from. `Mod` variants also have a clear advantage over monolithic (`Full`) models in bitext mining (both for LaBSE and mE5), even in the absence of explicit cross-lingual training (i.e., Mod$_{m}$). This suggests that multilingual training on `Full` results in negative interference (i.e., the curse of multilinguality), which is alleviated by our modular approach.

Cross-Lingual Training. Cross-lingual adapters, either trained on paraphrase data (Mod$_{mc\text{-}pp}$) or parallel data (Mod$_{mc\text{-}pl}$) can effectively reduce language bias and increase isomorphism of monolingual spaces (cf. Mod$_{m}$). Results further show that paraphrase- and parallel-CLA adapters benefit different types of cross-lingual tasks. On both LaBSE

	Monolingual tasks			Cross-lingual tasks					Alignement metrics		
	STS↑	**STR↑**	**CLS↑**	**STS↑**	**STR↑**	**CLS↑**	**Bitext Mining↓**		**Language Bias↓**		**RSIM↑**
Dataset	stsb	sick	sib	stsb	sick	sib	flores	tatoeba	stsb	sick	flores
mE5	72.5	74.2	74.0	54.1	61.0	73.5	1.85	9.89	23.22	12.11	0.60
Full_{en}	75.8	75.4	83.4	55.4	62.2	82.9	1.46	9.98	7.21	5.79	0.59
Full_{m}	**79.6**	**75.5**	85.5	60.2	64.1	85.2	0.62	7.85	2.60	3.16	0.67
Full_{c}	77.7	73.9	**85.6**	**66.7**	**67.7**	85.5	**0.26**	6.37	1.11	1.24	**0.74**
Full_{mc}	77.4	73.1	85.4	**66.7**	66.9	**86.5**	**0.26**	**6.33**	**1.05**	**1.14**	**0.74**
Mod_{en}	79.9	75.8	87.0	66.2	66.7	87.0	0.26	5.81	6.66	5.27	0.72
Mod_{m}	82.1	75.4	87.8	69.8	68.5	87.7	0.22	5.27	2.82	3.07	0.74
$\text{Mod}_{\text{mc-pp}}$	81.7	**76.4**	87.9	**73.2**	**70.5**	87.6	**0.20**	5.19	**1.58**	2.08	0.75
$\text{Mod}_{\text{mc-pl}}$	80.8	75.2	**88.5**	72.8	69.6	**89.0**	0.22	5.61	2.15	2.05	**0.82**
Mod_{mc}	**82.2**	**76.4**	88.3	73.0	70.0	88.1	**0.20**	**5.10**	1.63	**1.97**	0.78
Ablations											
Mod_{m}–LA	80.8	76.0	87.2	61.5	64.4	86.3	0.56	7.63	3.87	3.53	0.68
Mod_{c}	81.8	76.0	88.4	72.7	69.2	88.2	0.17	5.26	3.19	3.79	0.79

Table 2: Results of the mE5-based models for 10 languages. Reported results are averages over all languages in each evaluation dataset. The best result within the Full group and the Mod group on each dataset is denoted in **bold**. The second-best result in the Mod group is underlined.

and mE5, $\text{Mod}_{\text{mc-pl}}$ has the strongest performance in cross-lingual classification transfer (CLS), which correlates with the degree of isomorphism. In contrast, $\text{Mod}_{\text{mc-pp}}$ is better at STS/STR. This confirms the conflicting requirements of downstream tasks an importance of skill separation via modularity. Merging both CLA adapters in Mod_{mc}mitigates individual shortcomings, resulting in well-balanced performance across all tasks. Surprisingly, combining the two CLA adapters in Mod_{mc} improves the performance on monolingual tasks, despite each CLA adapter invidually reducing the monolingual STS/STR compared to Mod_{m}. Our complete Mod_{mc} setup thus makes the best use of our multi-parallel paraphrase dataset.

Ablation: Monolingual Specialization. Additional monolingual training for each language as an intermediate step before cross-lingual alignment distinguishes our modular approach from other popular MSE training strategies. We thus ablate the contribution of the monolingual specialization step (last two rows in Table 1 and Table 2). We first remove the LA step, i.e. we omit the MLM training with language-specific embedding layer and language adapter and directly train the monolingual SE adapter on the original MSE. For both LaBSE and mE5, this leads to a significant performance drop compared with Mod_{m}. Without language adaptation, adapter-based SE training even underperforms Full_{m}in monolingual tasks, but improves over it in cross-lingual tasks: this again suggest that modular multi-parallel monolingual SE training benefits cross-lingual semantic alignment more than multilingual training of all parameters. To isolate the contribution of the monolingual SE adapter, we remove the SE adapter for non-English languages from Mod_{mc} to get a Mod_{c} baseline: now the sentence encoding in other languages is learned only through the alignment to the English representations. We observe that while the task performance is only slightly affected, the language bias in STS/STR significantly increases, suggesting that the removal of monolingual SE training is very detrimental to the strong cross-lingual alignment of language-specific representation subspaces. The ablation results prove that our monolingual specialization steps are not only effective for improving monolingual performance of individual languages, but also plays an indispensable role in cross-lingual alignment.

6 Conclusion

Multilingual sentence encoders (MSE) encode sentences from many languages in a shared semantic spaces. As a consequence, they suffer from the curse of multilinguality and trade monolingual performance for cross-lingual alignment. Moreover, the choice of different types of data (paraphrases vs. parallel data) results in performance trade-offs across downstream tasks, In this work, we addressed these shortcomings via modularity: we (1) first specialize monolingual SEs via monolingual contrastive training on machine-translated paraphrase data, initializing them from the same MSE. Shared initialization and multi-parallel paraphrase data then facilitate the cross-lingual alignment of the monolingual SEs, which we are able to improve with lightweight cross-lingual alignment adapters. We show (i) that this modular approach yields gains w.r.t. both monolingual and cross-lingual performance and (ii) that MT can help train effective sentence encoders.

Limitations

We only experiment with encoder-based MSEs like LaBSE and mE5. Though this is the mainstream architecture for most MSEs, there are also pre-trained MSEs with the encoder-decoder architecture (Duquenne et al., 2023). Since the pre-training training objectives of such models are different from the encoder-based models we use (i.e. MLM language modelling and contrastive sentence embedding learning), our current modular training approach cannot be directly applied to them without adaptations. We thus leave the application of our modular approach to improve encoder-decoder MSEs to future work.

Having language-specific modules for each language requires that the language of the input text is known. If the language is unknown, a prior language identification step is needed to determine it, as we do not have a built-in language detection module. Fortunately, language identification is generally straightforward and reliable models that recognize hundreds of languages are readily available (Kargaran et al., 2023).

Ethics Statement

Our experiments use publicly available datasets and benchmarks for training and evaluation: these are all commonly used in the NLP research. No personal information or sensitive data are involved in our work. Existing biases in the public datasets, our machine-translated datasets and pre-trained models can still be relevant concerns, as we do not specifically mitigate them in this work.

References

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Vladimir Araujo, Andrés Carvallo, Souvik Kundu, José Cañete, Marcelo Mendoza, Robert E. Mercer, Felipe Bravo-Marquez, Marie-Francine Moens, and Alvaro Soto. 2022. Evaluation benchmarks for Spanish sentence representations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6024–6034, Marseille, France. European Language Resources Association.

Mikel Artetxe and Holger Schwenk. 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, Florence, Italy. Association for Computational Linguistics.

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

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. *CoRR*, abs/2401.10440.

Léo Bouscarrat, Antoine Bonnefoy, Thomas Peel, and Cécile Pereira. 2019. STRASS: A light and effective method for extractive summarization based on sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 243–252, Florence, Italy. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Ethan C. Chau and Noah A. Smith. 2021. Specializing multilingual language models: An empirical study. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 51–61, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

D. R. Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242.

Slawomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2020. Evaluation of sentence representations in Polish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1674–1680, Marseille, France. European Language Resources Association.

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

Konstantin Dobler and Gerard de Melo. 2023. FOCUS: Effective embedding initialization for monolingual specialization of multilingual models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13440–13454, Singapore. Association for Computational Linguistics.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. SONAR: sentence-level multimodal and language-agnostic representations. *CoRR*, abs/2308.11466.

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

Katja Filippova and Yasemin Altun. 2013. Overcoming the lack of parallel data in sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1481–1491, Seattle, Washington, USA. Association for Computational Linguistics.

Pengzhi Gao, Liwen Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2023. Learning multilingual sentence representations with cross-lingual consistency regularization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 243–262, Singapore. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Goran Glavaš and Ivan Vulić. 2020. Non-linear instance-based cross-lingual mapping for non-isomorphic embedding spaces. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7548–7555, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 422–430, Online. Association for Computational Linguistics.

Jiyeon Ham and Eun-Sol Kim. 2021. Semantic alignment with calibrated similarity for multilingual sentence embedding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1781–1791, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hebatallah A. Mohamed Hassan, Giuseppe Sansonetti, Fabio Gasparetti, Alessandro Micarelli, and Jöran Beel. 2019. Bert, elmo, USE and infersent sentence encoders: The panacea for research-paper recommendation? In *Proceedings of ACM RecSys 2019 Late-Breaking Results co-located with the 13th ACM Conference on Recommender Systems, RecSys 2019 Late-Breaking Results, Copenhagen, Denmark, September 16-20, 2019*, volume 2431 of *CEUR Workshop Proceedings*, pages 6–10. CEUR-WS.org.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matthew L. Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *CoRR*, abs/1705.00652.

Tomáš Hercig and Pavel Kral. 2021. Evaluation datasets for cross-lingual semantic textual similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 524–529, Held Online. INCOMA Ltd.

Christopher Hidey and Kathy McKeown. 2016. Identifying causal relations using parallel Wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1424–1433, Berlin, Germany. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Amir Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yuto Kuroda, Tomoyuki Kajiwara, Yuki Arase, and Takashi Ninomiya. 2022. Adversarial training on disentangling meaning and language representations for unsupervised quality estimation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5240–5245, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hauke Licht. 2023. Cross-lingual classification of political texts using multilingual sentence embeddings. *Political Analysis*, 31(3):366–379.

Meizhen Liu, Xu Guo, He Jiakai, Jianye Chen, Fengyu Zhou, and Siu Hui. 2023. InteMATs: Integrating granularity-specific multilingual adapters for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5035–5049, Singapore. Association for Computational Linguistics.

Zhuoyuan Mao, Prakhar Gupta, Chenhui Chu, Martin Jaggi, and Sadao Kurohashi. 2021. Lightweight cross-lingual sentence representation learning. 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 2902–2913, Online. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Isabelle Mohr, Markus Krimmel, Saba Sturua, Mohammad Kalim Akram, Andreas Koukounas, Michael Günther, Georgios Mastrapas, Vinit Ravishankar, Joan Fontanals Martínez, Feng Wang, Qi Liu, Ziniu Yu, Jie Fu, Saahil Ognawala, Susana Guzman, Bo Wang, Maximilian Werk, Nan Wang, and Han Xiao. 2024. Multi-task contrastive learning for 8192-token bilingual text embeddings. *CoRR*, abs/2402.17016.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Abinew Ali Ayele, Pavan Baswani, Meriem Beloucif, Chris Biemann, Sofia Bourhim, Christine de Kock, Genet Shanko Dekebo, Oumaima Hourrane, Gopichand Kanumolu, Lokesh Madasu, Samuel Rutunda, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Hailegnaw Getaneh Tilaye, Krishnapriya Vishnubhotla, Genta Indra Winata, Seid Muhie Yimam, and Saif M. Mohammad. 2024. Semrel2024: A collection of semantic textual relatedness datasets for 14 languages. *CoRR*, abs/2402.08638.

Marinela Parović, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2022. BAD-X: Bilingual adapters improve zero-shot cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1791–1799, Seattle, United States. Association for Computational Linguistics.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019.

Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy. Association for Computational Linguistics.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. Adapters: A unified library for parameter-efficient and modular transfer learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 149–160, Singapore. Association for Computational Linguistics.

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

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. LAReQA: Language-agnostic answer retrieval from a multilingual pool. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5919–5930, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Lütfi Kerem Senel, Benedikt Ebing, Konul Baghirova, Hinrich Schuetze, and Goran Glavaš. 2024. Kardeş-NLU: Transfer to low-resource languages with the help of a high-resource cousin – a benchmark and evaluation for Turkic languages. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1672–1688, St. Julian's, Malta. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. 2021. Language-agnostic representation from multilingual sentence encoders for cross-lingual similarity estimation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7764–7774, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3178–3192, Online. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. *CoRR*, abs/2402.05672.

Yaushian Wang, Ashley Wu, and Graham Neubig. 2022. English contrastive learning can learn universal cross-lingual sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9122–9133, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Gijs Wijnholds and Michael Moortgat. 2021. SICK-NL: A dataset for Dutch natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1474–1479, Online. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 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, Online. Association for Computational Linguistics.

Ziyi Yang, Yinfei Yang, Daniel Cer, and Eric Darve. 2021. A simple and effective method to eliminate the self language bias in multilingual representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5825–5832, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chihiro Yano, Akihiko Fukuchi, Shoko Fukasawa, Hideyuki Tachibana, and Yotaro Watanabe. 2024. Multilingual sentence-t5: Scalable sentence encoders for multilingual applications. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11849–11858, Torino, Italia. ELRA and ICCL.

Kaiyan Zhao, Qiyu Wu, Xin-Qiang Cai, and Yoshimasa Tsuruoka. 2024. Leveraging multi-lingual positive instances in contrastive learning to improve sentence embedding. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 976–991, St. Julian's, Malta. Association for Computational Linguistics.

A Languages

Table 3 lists the languages with their codes and scripts.

Code	Language	Script
am	Amharic	Ge'ez
ar	Arabic	Arabic
az	Azerbaijani	Latin
cs	Czech	Latin
de	German	Latin
en	English	Latin
es	Spanish	Latin
fr	French	Latin
ha	Hausa	Latin
it	Italish	Latin
kk	Kazakh	Cyrillic
ko	Korean	Hangul
ky	Kyrgyz	Cyrillic
mr	Marathi	Devanagari
nl	Dutch	Latin
pl	Polish	Latin
ru	Russian	Cyrillic
rw	Kinyarwanda	Latin
te	Telugu	Ge'ez
tr	Turkish	Latin
ug	Uyghur	Arabic
uz	Uzbek	Latin
zh	Chinese	Han (simplified)

Table 3: Languages with their code used in this paper and the scripts.

B Implementation Details

The pre-trained models and libraries used in our experiments are listed in Table 4. They are used only for research purposes in this work. We do not do specific hyperparameter tuning because of the large-scale MLM training and the robustness of contrastive learning against hyperparameters (Wang et al., 2022). Thus, we mainly use hyperparameters recommended by the previous work or default settings in the packages.

B.1 Full-Parameter Baselines

Both monolingual and cross-lingual contrastive learning on all baselines are done with a sequence length of 128, batch size of 128 and learning rate of 2e-5. To make a fair comparison with the modular variants, we train Full_{m} and Full_{c} for 3 epochs on the 600K monolingual or cross-lingual paraphrase data, respectively, while the Full_{mc} is obtained by 3 epochs of monolingual training followed by another 3 epochs of cross-lingual training. We found that further increasing the number of epochs will not improve the performance.

B.2 Modular Training

The parameter size of each module and training time for each step is reported in Table 5.

FOCUS The training of language-specific tokenizers and the initialization of language-specific embedding matrices is done using the `deepfocus`

Model	HuggingFace Name	License
LaBSE	sentence-transformers/LaBSE	apache-2.0
NLLB	facebook/nllb-200-3.3B	cc-by-nc-4.0
mE5 base	intfloat/multilingual-e5-base	mit
Libarary	**GitHub Link**	**License**
transformers	https://github.com/huggingface/transformers	apache-2.0
sentence-transformers	https://github.com/UKPLab/sentence-transformers	apache-2.0
adapters	https://github.com/adapter-hub/adapters	apache-2.0
deepfocus	https://github.com/konstantinjdobler/focus	mit

Table 4: Models and libraries used in our experiments.

Step	Module	Parameters %	Time
language adaptation	embedding layer, language adapter	8.4%	20h
sentence encoding	SE adapter	0.3%	20m
cross-Lingual alignment	CLA adapter	1.5%	20m * 2

Table 5: Parameter size (percentage of the original MSE size of 472 Million) and training time for each module. The training is done on a A100 40G GPU.

package (Table 4). We set the vocabulary size to 50K for each language. The dimensionality of fastText embeddings used to calculate token similarity is set to 300 as recommended. Other parameters remain as default. We use up to 10M sentences for the training of the tokenizer and the auxiliary fastText embeddings on each language.

Language Adaptation As the language adapter, we use a LoRA adapter (Hu et al., 2022) on key, query, value matrices of the attention layers, with a rank of 8, alpha of 16 and 0.1 dropout. For each language, we train the embedding layer and the language adapter for 200K steps, with a batch size of 128. For high-resource languages, 200K steps of training only cover a small portion of the available data in MADLAD-400 (Kudugunta et al., 2023). For low-resource languages, we use all data of the corresponding language from CC100 (Conneau et al., 2020) and MADLAD-400 (Kudugunta et al., 2023).

Monolingual SE Adapter As the monolingual SE training, we use a LoRA adapter (Hu et al., 2022) on all linear layers, with a rank of 8, alpha of 16 and 0.1 dropout. We use the 600K paraphrase data in the corresponding language for contrastive sentence embedding training for each language, with a sequence length of 128, batch size of 128 and learning rate of 2e-5 for 1 epoch in mixed precision.

Cross-Lingual Alignment For the training of CLA adapters, we use 600K bilingual paraphrase data as explained in section 3. Each adapter is trained with a sequence length of 128, batch size of 256 and learning rate of 2e-5 for 1 epoch in mixed precision. We use the parallel adapter (He et al., 2022) with default settings in `Adapters` (Poth et al., 2023) for CLA training.

C Datasets

We provide detailed information on the training and evaluation datasets. The datasets are used only for research purposes in this work.

C.1 Paraphrase Data

Table 6 provides an overview of the paraphrase datasets. The XNLI dataset is licensed with cc-by-nc-4.0. For the sources of other datasets, please refer to the information page[8].

C.2 STS/STR Evaluation Data

We use the test split of the datasets for zero-shot evaluation. In the following, we list the sources of STS/STR data for all individual languages and language pairs. Note that for symmetric pairs (e.g. en-de and de-en), the score in our experiments is the average of both directions.

STS17 The data for en, ar, es, en-ar, en-tr and es-en in the extended STS17 comes from the original STS17 (Cer et al., 2017). The data for de, fr, cs, de-en, en-fr, en-cs, cs-en, de-fr, fr-de, cs-de, de-cs, cs-fr and fr-cs is created by Hercig and Kral (2021). And the en-de, fr-en, nl-en and it-en data is translated by Reimers and Gurevych (2020). Through combining the data from Hercig and Kral (2021) and Reimers and Gurevych (2019), we get evaluation sets for nl-de, nl-fr, nl-cs, it-de, it-fr and it-cs.

[8] https://huggingface.co/datasets/sentence-transformers/embedding-training-data

Dataset	Description	Size
MNLI/XNLI	Multi-Genre NLI data. We build 128K (Anchor, Entailment, Contradiction) triplets using the original data.	128K
Sentence Compression	Pairs (long_text, compressed_text) from news articles.	108K
Simple Wiki	Matched pairs (English_Wikipedia, Simple_English_Wikipedia)	102K
Altlex	Matched pairs (English_Wikipedia, Simple_English_Wikipedia)	113K
Quora Duplicate Questions	Duplicate question pairs from Quora. We use the "triplet" subset.	102K

Table 6: Overview of paraphrase datasets. Except for XNLI, all of them are English datasets and are machine-translated into our target languages for training.

All data except for ko are from the SNLI domain, containing 250 sentence pairs per language pair. The ko data is translated from the English STS benchmark (Cer et al., 2017) by Ham et al. (2020), containing 2850 pairs in various domains. Results for en-cs, de-fr, cs-de, and cs-fr are calculated as the average of symmetric language pairs (e.g. de-fr is the average of de-fr and fr-de).

STSB Senel et al. (2024) translate the en data from the STS benchmark (Cer et al., 2017) into 5 Turkic languages: az, kk, ky, ug and uz. There are 800 test sentence pairs from various domains for each language. Since the other training data for Uyghur is written in the Arabic script, we transliterate the Cyrillic Uyghur data in the benchmark into the Arabic script using the Uyghur Multi-Script Converter[9]. The Turkic language data are combined with the dataset for ko (Ham et al., 2020) to form evaluation dat for ko-en, ko-az, ko-ky, ko-ug and ko-uz. For STSB, all cross-lingual results are the average of symmetric language pairs (e.g. az-kk is the average of az-kk and kk-az).

SICK We use the SICK dataset in English (Marelli et al., 2014), Polish (Dadas et al., 2020), Dutch (Wijnholds and Moortgat, 2021) and Spanish (Araujo et al., 2022) and combine them to create cross-lingual evaluation data for en-pl, en-nl, en-es, pl-nl, pl-es and nl-es. The test set size is 4.91K for each language (pair). All cross-lingual results are the average of symmetric language pairs.

STR24 We use the test data of the supervised track of STR24, including monolingual data for en (2600 pairs), am (342), ha (1206), rw (444), mr (298), te (297). We do not include Spanish because the public test set is not available, nor the Moroccan Arabic and Algerian Arabic because they are not supported by LaBSE. The data is curated primarily from news (Ousidhoum et al., 2024).

[9] https://github.com/neouyghur/Uyghur-Multi-Script-Converter