EnerGIZAr: Leveraging GIZA++ for Effective Tokenizer Initialization

Anonymous ACL submission

Abstract

Continual pre-training has long been considered the default strategy for adapting models to non-English languages, but struggles with initializing new embeddings, particularly for non-Latin scripts. In this work, we propose EnerGIZAr, a novel methodology that improves continual pre-training by leveraging statistical word alignment techniques. Our approach utilizes GIZA++ to construct a subword-level alignment matrix between source (English) and target language tokens. This matrix enables informed initialization of target tokenizer embeddings, which provides a more effective starting point for adaptation. We evaluate EnerGIZAr against state-of-the-art initialization strategies such as OFA and FOCUS across four typologically diverse languages: Hindi, Basque, Arabic and Korean. Experimental results on key NLP tasks – including POS tagging, Sentiment Analysis, NLI, and NER – demonstrate that EnerGIZAr achieves superior monolingual performance while also out-performing all methods for cross-lingual transfer when tested on XNLI. With EnerGIZAr[1], we propose an intuitive, explainable as well as state-of-the-art initialisation technique for continual pre-training of English models.

1 Introduction

As research into developing better and larger language models progresses, models for medium- and low-resourced languages continue to lag behind. English models are always the first to be introduced to new developments in LLM pre-training. Sometimes major advancements also include multilingual models as a secondary checkpoint, but this is seldom the case. This leaves researchers working on non-English languages with two primary options. First, to train their own models with the technological enhancements proposed. This option comes with restrictions on data sizes and available compute. Newer methodologies often use large unstructured English corpora like C4 (Raffel et al., 2020), OSCAR (Ortiz Suárez et al., 2019), OpenWebText (Gokaslan et al., 2019), etc. However, the corpora available for other languages are hardly comparable in size to the unstructured English datasets, and therefore the resulting models are often sub-optimal. Wu and Dredze (2020) showed that monolingual models trained for low-resource languages performed significantly worse than mBERT despite mBERT having very limited representations for low-resource languages.

The second option, and the more commonly used one, is to adapt existing English or multilingual models to a particular target language. This option is preferable for reasons such as computational efficiency, low data requirements, etc. The most commonly accepted methodology in practice for achieving this has been continual pre-training. Continual pre-training uses an English or multilingual model as a checkpoint and continues training for the pre-training objective. Continual pre-training decidedly results in a better model and is more efficient, however, it does come with its restrictions. Since an English or multilingual model is used as the starting checkpoint, it forces one to use the vocabulary of the source model, which might not be fit for several languages, especially those in non-Latin scripts. Even when using a multilingual vocabulary, research by Rust et al. (2021) has shown that the representation of most medium- and low-resourced languages is meek at best.

This bottleneck has led to significant work in optimally initializing new tokenizer entries (Wang et al., 2019; Tai et al., 2020; Hewitt, 2021) or adapting models to target language tokenizers (Minixhofer et al., 2022; Dobler and de Melo, 2023; Liu et al., 2024). The challenge arises in finding a methodology that can consistently initialize new embeddings with minimal supervision across hundreds of languages with varying scripts and other

[1] Anonymized Github link

typological factors.

In this research we propose EnerGIZAr, a methodology for improved continual pre-training. We show that by tokenizing parallel corpora, followed by aligning them with GIZA++ (see example in Figure 1), a statistical alignment tool trained using Expectation Maximisation (EM), we can initialize an alignment matrix between the source (English-only) and target language tokens. Using said alignment matrix to initialize target tokenizer embeddings results in an excellent initial checkpoint for continual pre-training. Moreover, having a subword-to-subword alignment matrix makes the methodology particularly transparent and interpretable allowing for the possibility of manual or semi-automated modifications to the matrix to further enhance the initialization. We attempt to answer the following main research questions in this work:

- Is it feasible to initialize a model in a target language using parallel data and a word alignment tool?
- Can this initialized model, when continually trained, compete with other SOTA initialisation strategies like OFA (Liu et al., 2024) and FOCUS (Dobler and de Melo, 2023) for monolingual performance?
- Which of the initialisation strategies – SOTA versus EngerGIZAr – preserves the most cross-lingual signals, therefore resulting in the best model for cross-lingual use cases?

We perform experiments on four languages (Hindi, Basque, Arabic, and Korean) with widely differing typological features, such as script, geolocation and morphology. We evaluate all models on downstream tasks with real-world use cases, including part-of-speech tagging, sentiment analysis, natural language inference, and named entity recognition. We also test all the methods' cross-lingual performance on the XNLI (Conneau et al., 2018) dataset. Our results illustrate that EnerGIZAr outperforms continual pre-training baselines as well as SOTA initialisation methods FOCUS and OFA, both in purely monolingual as well as cross-lingual testing.

The remainder of this paper is organised as follows. We cover related work in Section 2, with 2.1 covering related embedding initialisation strategies while 2.2 covers statistical alignment methods. Section 3 covers the methodology and formulation of the work, while Section 4 details the experimental protocol, hyperparameters, data, models, and additional information. Finally, Section 5 details the results of all experiments including monolingual and cross-lingual settings.

2 Related Work

2.1 Embedding Initialisation Strategies

Continual pre-training, the most commonly used practice to adapt pre-trained models to new domains and languages simply uses all the components of a transformer and continues training for MLM objectives with additional monolingual data (Gururangan et al., 2020). A better alternative to this can be continual pre-training using a tokenizer in the target language to better adapt to the vocabulary of the target language (Minixhofer et al., 2022). However, in this case, the embedding layer from the source model is completely discarded for a new randomly initialized embedding layer for the target language tokens. Although training an embedding layer from scratch increases convergence time, it usually results in a better model than vanilla continual pre-training, given sufficient data. Another advantage is the reduced length of tokenized text passed to the model since this allows the model to process more information in a single pass.

However, rather than random initialization, recent developments have proposed ideas for a more informed initialization of target language embeddings. WECHSEL (Minixhofer et al., 2022), FOCUS (Dobler and de Melo, 2023) and OFA (Liu et al., 2024) all rely on multilingual static word embeddings in a shared space as auxiliary embeddings. These methods enhance the transfer of embeddings by incorporating information from a static embedding space, such as FastText (Mikolov et al., 2018).

The WECHSEL method (Minixhofer et al., 2022) focuses on efficient initialization of subword embeddings by utilizing bilingual dictionaries to enhance knowledge transfer between languages. A shared static embedding space, aligned with the help of a bilingual dictionary, is used to compute the similarity between source and target sub-words. Next, a set of k-nearest neighbours in the source language is used to initialize target sub-words. However, it relies heavily on the quality and availability of the bilingual dictionaries as well the static embeddings used for alignment.

FOCUS (Dobler and de Melo, 2023), which stands for Fast Overlapping Token Combinations

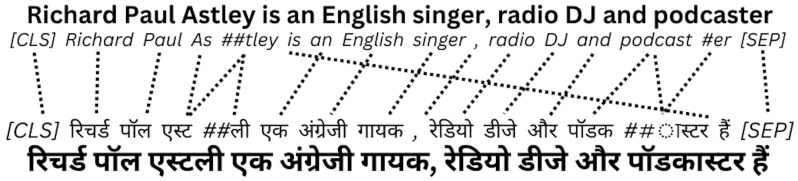

Figure 1: An example of two parallel sentences in English (above) and Hindi (below) and their tokenized forms (using *bert-base-cased* for English and *hindi-bert* for Hindi), aligned using GIZA++.

Using Sparsemax, is an innovative method for adapting pre-trained models to low-resource languages. The core idea behind FOCUS is to represent newly added tokens in a vocabulary as combinations of overlapping tokens found in the source vocabulary. This overlap is determined based on semantic similarity in an auxiliary token embedding space (FastText). The similarity computer between source and target tokens is converted to weights using SparseMax (Martins and Astudillo, 2016), and the weights are subsequently used for initialising a target word with the k-nearest neighbours.

The OFA (Liu et al., 2024) framework employs factorized embeddings to optimize computational efficiency while ensuring robust model performance. By dividing embeddings into language-agnostic and language-specific components, OFA reduces the number of parameters needed for training, leading to faster convergence and a lower environmental impact during pre-training. OFA uses ColexNet+ (Liu et al., 2023) embeddings as its source of multilingual information, creating a bipartite graph using a fixed set of neighbours for each target sub-word. Esentially, OFA builds on the works of WECHSEL and FOCUS, introducing the factorisation component and replacing the source of static embeddings with ColexNet+, which is more conceptually grounded and potentially a better source of cross-lingual signals. Once more, however the quality of the static embeddings heavily determines the initialisation.

In our work, we present a different take on embedding initialisation by bypassing the need of pre-trained multilingual embeddings, or orthogonal embedding alignment techniques used in previous work. Instead, we rely on statistical sub-word alignment. Our work can be related to Rémy et al. (2024), which was tested on the Mistral-family of models for Dutch and Tartu, where parallel data along with FastAlign was used to find the nearest neighbours for a target sub-word to be newly initialized. The key differences being the use of FastAlign, which prioritizes speed and efficiency over deep probabilistic modelling in contrast to GIZA++, the lack of alternate initialisation strategies like direct copying and random normal initialisation, as well as filtering and refinement strategies for the alignment matrix. In addition, our method is also thoroughly evaluated for an extensive set of tasks and languages in comparison with the SOTA of FOCUS and OFA.

Our approach starts from the intuition that SMT-based word alignment provides a more raw source of information overlap between two vocabularies, even though embedding similarities may be more granular and contain more information. Different from previous work, we also hypothesize that working with an alignment matrix - rather than potential nearest neighbours - allows for a more exhaustive solution, since each target sub-word can be influenced by each source sub-word independently and without constraints due to the k hyper-parameter for nearest neighbours.

2.2 Statistical Word Alignment Tools

Word alignment tools, such as GIZA and its successor GIZA++ (Och and Ney, 2003), run on large chunks of parallel data and have played a crucial role in NLP by facilitating the identification of translational equivalence between words. While NMT tools like LaBSE (Wang et al., 2022) may slightly eclipse SMT tools in performance, SMT tools still remain more efficient, explainable, and transparent, which is one of our motivations for using GIZA++ in this research.

GIZA++ is one of the most widely used tools for statistical word alignment, implementing IBM models (Brown et al., 1993) for word alignment tasks, allowing for the extraction of alignment probabilities between words. It operates by running alignments in both directions, i.e., source to target and target to source, and then combines the results to improve the quality of alignments. This process ensures that only one-to-one alignments are re-

tained in the final output, thereby increasing precision. Its features include the implementation of the full IBM-4 and IBM-5 alignment models, as well as the Hidden Markov Model (HMM). GIZA++ implements several key features that distinguish it from its predecessor, GIZA. The HMM implementation includes techniques such as Baum-Welch training and the Forward-Backward algorithm, and it also applies various smoothing methods for parameters related to fertility and distortion, which helps in refining the alignment results.

FastAlign (Dyer et al., 2013), designed for speed and efficiency in word alignment tasks, utilizes a simplified version of the GIZA++ algorithm, using only the IBM Model-2. It is less precise compared to GIZA++, especially for highly reordered languages, but since it allows faster processing, it is often more suitable for real-time applications.

3 Methodology

We begin by defining the problem mathematically. Let F^s be a source transformer with its usual components: tokenizer, Tok^s for vocabulary, W^s, embedding layer, Emb^s of size $len(W^s) \times 768$ and the subsequent encoder Enc^s. Our goal, given a monolingual corpus in a target language (M^t) and a source of cross-lingual signals, is to arrive at the best possible target transformer F_t. While methods like WECHSEL, FOCUS and OFA have attempted to use multilingual static embeddings as their source of cross-lingual signals, we use parallel data as our cross-lingual signal.

To detect corresponding sub-words between the source and target language, we train GIZA++ on our parallel corpus. The default training pipeline runs five iterations each of IBM Model 1, HMM, Model 3, and Model 4. Model 1 uses word translation probabilities ($p(y|x)$, where x is a source language word and y is a target language word) for learning alignments. HMM and Model 4 rely on relative reordering, while Model 3 uses a fertility model. For our work, we only use Model 4 for alignment, to make the entire pipeline significantly more efficient and cut down alignment times by up to 300%. We use the $grow - diag - final - and$ heuristic for alignment, which considers alignments from both directions, i.e., source-target as well as target-source.

Given the parallel corpus $P^{s,t}$, we first tokenize the respective parts P^s using the tokenizer Tok^s and P^t using the target tokenizer Tok^t to obtain the sub-word tokenized parallel corpora $P^{s,t}_{tok}$. We then use the sub-word tokenized parallel corpora with IBM Model 4 to train and run an alignment model on the tokenized parallel data. This results in a translation probability dictionary which can be represented as a matrix $D^{t,s}$. This matrix indicates the probability a source sub-word x can be translated into a target sub-word y as $p(y|x)$.

$$P^t_{tok} = Tok_t(P^t),\ P^s_{tok} = Tok_s(P^s) \quad (1)$$

$$D^{t,s}(P^t_{tok}, P^s_{tok}) = [p(y|x)]\ \forall\ y \in W_t,\ x \in W_s \quad (2)$$

The following post-processing is applied to the matrix $D^{s,t}$. First, probabilities below a hyperparameter δ are set to 0. Furthermore, if the probabilities for a target word y are too widely distributed, i.e., all probabilities $0 \geq p_{y,x} \forall x \in W_s \leq 0.1$, we initialize the word's embedding using a normal distribution centered at the mean of all source embeddings, Emb_s. Finally, we also find source and target sub-words that are identical (numbers, special characters and reserved tokens) and explicitly set the probabilities of these matching words to 1.0 while setting all other probabilities for the source word to 0.0, thus effectively ensuring identical sub-words are not newly initialized.

Finally, to initialize the target tokenizer embeddings Emb^t, we simply use the cross product,

$$Emb^t = D^{t,s} \times Emb^s \quad (3)$$

Essentially, each target sub-token embedding is initialized as a weighted average of all relevant source sub-tokens embeddings. With the newly initialized Emb^t the encoder Enc^s can be trained for MLM with target language data M^t, while using the appropriate tokenizer Tok^t. An overview of the proposed methodology is presented by Figure 2.

4 Experimental Setup

We perform experiments for a set of four languages: Hindi, Basque, Korean, and Arabic. The languages were selected based on diversity in scripts, geolocation and language families. Table 1 presents an overview of the resources for each target language.

4.1 Pre-training

For each language *bert-base-cased* was used as the starting monolingual model F^s. As can be

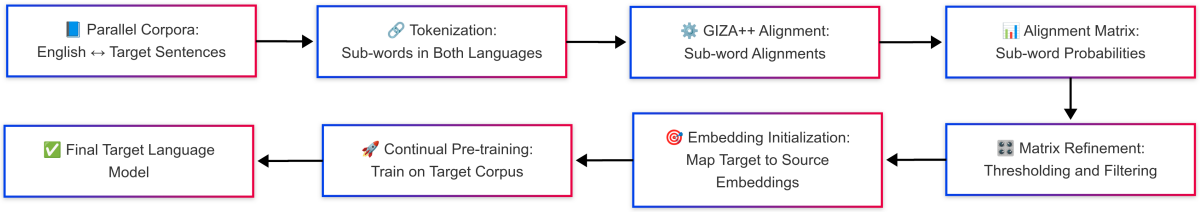

Figure 2: A summary of the proposed methodology of EnerGIZAr.

Language	Wiki (tokens)	OpusMT (tokens)	Tokenizer	Tasks
Hindi	42.10 M	7.29 M	Hindi-BERT	UDPOS, Sentiment, Topic
Basque	69.01 M	7.15 M	BertEUS	UDPOS, Sentiment, Topic
Arabic	278.60 M	8.58 M	CAMeLBERT-msa	NER, Stance, Emotion
Korean	133.66 M	5.17 M	KorBERT	NER, NLI, Topic

Table 1: Overview of the target languages used for the experiments, their available resources – both monolingual (Wiki) and parallel (OpusMT) – the target Tokenizer used for the transfer, and downstream tasks used for testing.

observed from Table 1, all languages can be considered medium-resourced based on the available monolingual and parallel corpora sizes. For each language, their respective Wikipedia dump was used as the monolingual resource M^t, M^s, and Opus-MT as the source of all parallel data $P^{s,t}$. Wikipedia was chosen over CommonCrawl, C4 or OSCAR as it significantly reduces the duration of experimentation, allowing us to iterate and tune parameters such as δ. For example, the Wiki size of Hindi (see Table 1) is approximately 42.10 million tokens, whereas the size of Common Crawl for Hindi is approximately 1.8 billion tokens – roughly 40 times larger. While models trained on the Common Crawl dump would undoubtedly result in better overall target language models, the experimentation time for each language would be 40-50 times slower. Moreover, reducing the amount of pre-training data helps us to better simulate a lower-resource setting.

As the source language tokenizer Tok^s we used the standard tokenizer of *bert-base-cased*, while as Tok^t we used the appropriate tokenizer from the baseline monolingual models available to streamline comparison with the respective models. We used *Hindi-BERT*[2], *BERTeus*[3], *KR-Medium*[4] and *CAMeLBERT*[5] for Hindi, Basque, Korean and Arabic, respectively.

To align the tokenized English and target language instances, we use GIZA++ with IBM Model 4, with $grow - diag - final - and$ as the symmetrization heuristic, maximum fertility of 10 and maximum sentence length of 101. After obtaining the matrix $D^{t,s}$, we apply the post-processing as described in the previous section. Based on preliminary experimentation on Hindi we found a δ of 0.1 to be the best-performing; however, this could vary slightly depending on the language and the tokenizer sizes. For the continual training, we train with M^t, with early stopping, with a learning rate of $1e-4$, and a maximum sequence length of 512.

For adequate comparison with the state of the art, we also train our own OFA and FOCUS models using the identical resources described for EnerGIZAr for the languages under consideration, and by relying on the official codebase of each project. For OFA, we used the OFA-768 models since these are, in terms of parameters, identical to the other models with which they are being compared. This is a significant contribution, as FOCUS and OFA-768 are both state-of-the-art embedding initialization methods which have not been directly compared before. For cursory testing for the pre-trained models, we evaluate for the standard Masked Language Modelling (MLM) loss on a held-out validation set for the target language. Since the tokenizers for each target language model are identical (Tok^t), the MLM loss should be directly comparable. We define the MLM loss as,

$$L_{mlm}(x_i) = -logP(x_i|h_i^L) \quad (4)$$

where for a single masked token x_i the loss is calculated as the cross-entropy between x_i and h_i^L, where h_i^L is the output vector from the last transformer layer (L) for each masked token i.

[2]https://huggingface.co/monsoon-nlp/hindi-bert
[3]https://huggingface.co/berteus-base-cased
[4]https://huggingface.co/snunlp/KR-Medium
[5]https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-msa

4.2 Downstream Testing

For each language we target 3 varied tasks for downstream testing, covering a wide range of language understanding, from syntactic tasks like Part-of-Speech (POS) tagging and Named Entity Recognition (NER) to affect-based subjective tasks like Sentiment and Stance Detection, as well as reasoning-based tasks like Natural Language Inference (NLI). To this purpose we referred to each language's respective standard language evaluation benchmarks i.e., Indic-GLUE (Kakwani et al., 2020) for Hindi, BasqueGLUE (Urbizu et al., 2022) for Basque, ALUE (Seelawi et al., 2021) for Arabic and KLUE (Park et al., 2021) for Korean.

Tasks were considered as long as sufficient training data was available (some tasks had less than 1000 training samples available and were therefore not considered). We use the Universal Dependencies (Nivre et al., 2017) project for the POS data (HDTB Treebank For Hindi, BDT Treebank for Basque). For Sentiment Detection in Hindi we use the IITP-PR (Akhtar et al., 2016) dataset, while for topic classification we use the WSTP (Wikipedia Section Title Prediction) dataset formulated as a Multiple Choice Question Answering Task. For Sentiment Detection in Basque, we use the BEC dataset (Agerri et al., 2020), while for topic classification we use the BHTC dataset from the same benchmark. For Arabic, we use the popular WikiANN (Rahimi et al., 2019) dataset for NER, for Stance detection we use the ANS-stance dataset (Khouja, 2020), and for multi-label Emotion we use the Arabic subset from the SemEval 2018 Task 1 data (Mohammad et al., 2018). Finally, for Korean, all 3 tasks, NER, NLI & Topic Classification were introduced in the KLUE benchmark (Park et al., 2021).

For each downstream task we use the provided validation and test splits. We perform model selection on the validation set to pick the best model. All tasks were trained with an initial learning rate of 5e-5 with a weight decay of 0.01 with around 10% of the total steps being used for warmup. We run each experiment 3 times and report the mean performance along with the standard deviation. For comparisons with the current SOTA, we also evaluate the downstream tasks for the OFA-768 and FOCUS models trained in the previous setup. For each language, we also test with the original monolingual BERT model whose tokenizer we use as Tok^t for the embedding transfer.

4.3 Cross-lingual Testing

An often used feature of multilingual models is their capacity for cross-lingual transfer. Barring availability of annotated data in the target language, an English (or other high-resource language) dataset can be used to train the model while inferring on the target language. While not as effective as training on the same language, cross-lingual transfer has proven an excellent alternative for non-English languages for which hardly any annotated data is available. In order to evaluate the cross-lingual capabilities of our approach, we also perform basic cross-lingual testing. We use the popular XNLI (Conneau et al., 2018) dataset to this purpose, training each model in English with 40,000 samples from the training set, while testing it for the 4 target languages under consideration. For consistency, we used the same settings as described in the previous section.

5 Results

5.1 Pre-training

We evaluate the effectiveness of pre-training using MLM loss on a held-out validation set. Figure 4 shows the validation loss for the models trained for *Hindi*. The baseline model represents the bert-base-cased model, an English-only model, continually pre-trained by using the tokenizer Tok_t from hindi-bert[6]. The *OFA-768* and *FOCUS* models represent the respective state-of-the-art models described in the previous section, initialized for Hindi using the given Tok_t. From the figure it is evident that *EnerGIZAr* not only initialized a better starting model but resulted in a better final model post continual training. The validation loss plots for the other languages show similar trends (see Appendix A).

Before looking at the results of downstream testing, we first examine the initializations made by each method. All tested initialisation methods follow three stages. First, identical or similar tokens are directly initialised from their source counterparts. Next, the respective methodology is applied, i.e., using the OFA multilingual static embeddings in the case of *OFA*, the FastText auxiliary embeddings in the case of *FOCUS* and the GIZA++ alignment using parallel data in the case of *EnerGIZAr*. Finally, embeddings identified as poor quality during the main initialisation step, are initialised using a normal distribution centred at the mean of

[6]https://huggingface.co/monsoon-nlp/hindi-bert

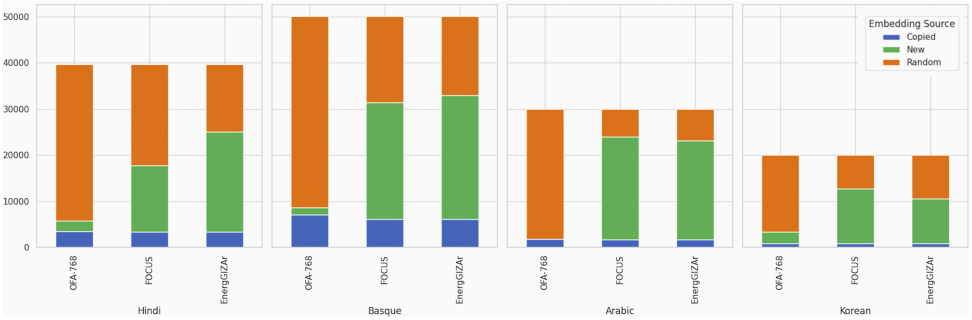

Figure 3: An overview of the different types of initialisations, i.e., Copied, New, and Random, performed by OFA, FOCUS and EnerGIZAr, on all 4 tested languages (Hindi, Basque, Arabic, Korean).

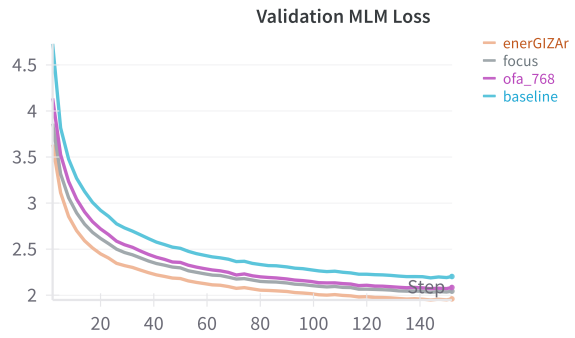

Figure 4: Illustration of the validation MLM loss for the baseline, OFA, FOCUS and our EnerGIZAr models.

all source embeddings. In most cases, copied embeddings are expected to have the highest quality, newly initialized the next best quality, and randomly initialized embeddings the lowest.

Figure 3 shows the results of the analysis. For Hindi and Basque, *EnerGIZAr* initialises the largest amount of new embeddings, followed by FOCUS. However, for Arabic and Korean, FOCUS initialises the highest amount of new embeddings while *EnerGIZAr* does second best. OFA always initialises the lowest amount of new embeddings. Among all languages, Basque had the highest initialisations by copying, potentially because this language shares the Latin script with English, while Korean had the fewest copied embeddings, resulting in more than 90% of the embeddings being newly or randomly initialised for all methods.

5.2 Downstream Testing

The results of the tests on all downstream tasks (measured in Macro-F1) are provided in Table 2. All methods result in an identical model in terms of parameters and architecture, allowing a fair comparison. From the results, it is evident that *EnerGIZAr* consistently outperforms the continual pre-training monolingual baseline and both state-of-the-art initialization methods, OFA and FOCUS, with only one exception: for Part-of-Speech tagging for Basque, OFA leads to the best result. The performance difference between OFA, FOCUS, and *EnerGIZAr* is minimal but consistent across all languages and tasks. Due to the closeness of these results as well as the overlap of the standard deviations we performed a one-tailed paired t-test first between FOCUS and EnerGIZAr with N=*36* (3 seeds, 3 tasks, 4 languages) to test statistical significance. We find that the results are extremely significant with a p-value of 0.0002 with a 95% confidence interval of *0.379* to *1.112*, with a mean difference of +0.7460. We perform a second one-tailed paired t-test between OFA-768 and EnerGIZAr. The outcome was identical with a p value of 0.0003, making the results statistically significant. The mean difference was even larger with +0.8572 with a 95% confidence interval of *0.4269* to *1.2875*. This makes *EnerGIZAr* the state-of-the-art method for embedding initialization for continual pre-training for monolingual use cases, with FOCUS being the second-best option in most scenarios.

Regarding the tasks itself we observe that the improvement is more apparent for the more semantic tasks such as Sentiment Classification and Natural Language Inference (NLI), while it is minor for the more syntactically informed tasks such as NER and POS. Concerning the latter, we can argue that both POS and NER are highly mature tasks with limited potential for further significant advancements due to saturation. The language with the lowest noticeable improvements on the downstream tasks compared to the baseline and state-of-the-art models is Arabic. Looking back at Table 1, we see that Arabic was the language with the highest amount of available data for continual training. We thus

Method	Hindi			Basque		
	UDPOS	Sentiment	Topic	UDPOS	Sentiment	Topic
bert-base	97.28 ±0.02	72.57 ±1.98	79.20 ±0.35	95.49 ±0.10	69.40 ±0.28	57.20 ±1.84
OFA-768	97.37 ±0.05	75.61 ±0.78	80.95 ±0.37	**95.65 ± 0.08**	67.40 ±0.77	59.66 ±0.74
FOCUS	97.43 ±0.03	74.68 ±1.63	80.86 ±0.35	95.61 ±0.16	68.49 ±0.47	59.50 ±1.03
EnerGIZAr	**97.46 ± 0.04**	**76.08 ± 0.67**	**82.68 ± 0.21**	95.61 ± 0.07	**69.76 ± 0.47**	**60.15 ± 0.77**

Method	Arabic			Korean		
	NER	Stance	Emotion	NER	NLI	Topic
bert-base	90.21 ±0.25	68.91 ±1.58	58.64 ±2.43	80.64 ±1.20	71.44 ±0.63	83.70 ±0.60
OFA-768	91.04 ±0.77	68.70 ±1.41	61.91 ±0.58	81.74 ±0.47	73.30 ±0.83	82.94 ±0.30
FOCUS	91.08 ±0.34	69.30 ±1.99	61.77 ±0.44	81.18 ±1.04	73.67 ±0.55	84.02 ±0.26
EnerGIZAr	**91.58 ± 0.64**	**69.42 ± 1.96**	**62.08±0.54**	**81.87± 0.40**	**75.43± 0.45**	**84.75± 0.40**

Table 2: Results for downstream testing of Baseline, OFA-768, FOCUS and our EnerGIZAr models for Hindi and Basque (above), and Arabic and Korean (below).

	Hindi	Basque	Arabic	Korean
bert-base (LAPT)	42.61 ±0.39	46.00 ±2.22	44.00 ±1.14	45.36 ±0.52
OFA-768	54.74 ±0.50	55.21 ±2.58	51.68 ±2.01	52.95 ±0.29
FOCUS	55.29 ±0.83	54.27 ±2.50	51.96 ±2.05	47.51 ±0.99
EnerGIZAr	**59.36 ±0.76**	**58.19 ±2.08**	**52.52 ±1.34**	**53.41 ±0.70**

Table 3: Results for Cross-lingual testing with the XNLI benchmark (trained on English, tested on the four target languages) for the baseline as well as ofa-769 and FOCUS compared to EnerGIZAr.

hypothesize that the initialization might not have been that impactful, i.e., when there is sufficient pre-training data, the model will probably be able to better converge, irrespective of poor initialization.

5.3 Cross-lingual Testing

Table 3 shows the results of cross-lingual testing for all 4 target languages with every model. It is clear that EnerGIZAr surpasses other methods in terms of cross-lingual capabilities when applied to the task of NLI. We hypothesize that this is due to the direct source of cross-lingual signals grounded in the parallel data, in contrast to the multilingual embeddings used for the other methods which are a more indirect source of cross-lingual information. Moreover, the alignment matrix ensures that little to no information is lost for a sub-word, compared to nearest-neighbour approaches. Moreover, we observe that the difference is more pronounced for languages where we have lower amounts of pre-training data available, such as Hindi and Basque, whereas the gap is smaller for a language with more extensive pre-training data, such as Arabic.

6 Conclusion

We introduce a new embedding initialisation strategy, EnerGIZAr, which uses the statistical alignment tool GIZA++ along with parallel data to initialise embeddings for a target language given an English-only model. Through extensive experiments in both monolingual downstream tasks as well as cross-lingual testing, we demonstrate that our method outperforms standard baselines as well as state-of-the-art initialisation methods. While the results for monolingual testing are closer, requiring paired t-tests to confirm the superiority of EnerGIZAr, in cross-lingual testing, EnerGIZAr outperforms current SOTA methods with ease, making it the clear choice for cross-lingual deployment scenarios. Although EnerGIZAr requires small amounts of parallel data, it does not require pre-trained multilingual static embeddings or auxiliary embeddings in any form. This might not be a direct advantage since all methods discussed require the availability of some form of cross-lingual signals, however, the requirements for EnerGIZAr differ slightly which could be useful in certain scenarios where availability of embeddings is sparse. EnerGIZAr also offers more interpretability due to the transparency of the alignment matrix and GIZA++, in contrast to using pre-trained static embeddings for alignment which are relatively more opaque. While we have not yet explored this aspect of EnerGIZAr and have left it for future research, we wish to utilize the transparency aspect by performing selected manual edits to the alignment matrix. Additionaly, in future work, we aim to demonstrate the effectiveness of EnerGIZAr for decoder models and compare with the work of Remy et al (2024).

Limitations

While EnerGIZAr demonstrates strong improvements in embedding initialization for continual pre-training, several limitations must be acknowledged. Firstly, EnerGIZAr relies on the availability of high-quality parallel corpora for subword alignment using GIZA++. This dependence makes it less applicable to languages with extremely limited or nonexistent bilingual resources, potentially reducing its effectiveness in extremely low-resource scenarios. Secondly, while the study covers four typologically diverse languages (Hindi, Basque, Arabic, and Korean), further validation is needed for other language families, especially those with agglutinative or polysynthetic structures. The methodology may require adaptation to maintain its effectiveness across these linguistic typologies. Lastly, the current experiments focus on encoder-based models (e.g., BERT-like architectures). The effectiveness of EnerGIZAr for initializing embeddings in decoder-based models, such as GPT-style autoregressive transformers, remains unexplored and warrants further research especially considering the success of decoder-based models in recent times.

Acknowledgments

Anonymised for reviewing.

References

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for Basque. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France. European Language Resources Association.

Md Shad Akhtar, Ayush Kumar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. A hybrid deep learning architecture for sentiment analysis. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 482–493, Osaka, Japan. The COLING 2016 Organizing Committee.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 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.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. 2019. Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus`.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

John Hewitt. 2021. Initializing new word embeddings for pretrained language models.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Jude Khouja. 2020. Stance prediction and claim verification: An Arabic perspective. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, Seattle, USA. Association for Computational Linguistics.

Yihong Liu, Peiqin Lin, Mingyang Wang, and Hinrich Schuetze. 2024. OFA: A framework of initializing unseen subword embeddings for efficient large-scale multilingual continued pretraining. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1067–1097, Mexico City, Mexico. Association for Computational Linguistics.

Yihong Liu, Haotian Ye, Leonie Weissweiler, Philipp Wicke, Renhao Pei, Robert Zangenfeind, and Hinrich Schütze. 2023. A crosslingual investigation of conceptualization in 1335 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 12969–13000, Toronto, Canada. Association for Computational Linguistics.

André F. T. Martins and Ramón Fernandez Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. *CoRR*, abs/1602.02068.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

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.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jungwoo Ha, and Kyunghyun Cho. 2021. Klue: Korean language understanding evaluation. *Preprint*, arXiv:2105.09680.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

François Remy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. 2024. Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of llms for low-resource nlp. *Preprint*, arXiv:2408.04303.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. 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 3118–3135, Online. Association for Computational Linguistics.

Haitham Seelawi, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. 2021. ALUE: Arabic language understanding evaluation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 173–184, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. exBERT: Extending pre-trained models with domain-specific vocabulary under constrained training resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439, Online. Association for Computational Linguistics.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2022. BasqueGLUE: A natural language understanding benchmark for Basque. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612, Marseille, France. European Language Resources Association.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, and Dong Yu. 2019. Improving pre-trained multilingual model with vocabulary expansion. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 316–327, Hong Kong, China. Association for Computational Linguistics.

Weikang Wang, Guanhua Chen, Hanqing Wang, Yue Han, and Yun Chen. 2022. Multilingual sentence transformer as a multilingual word aligner. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2952–2963, Abu Dhabi, United

Arab Emirates. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

A Validation Loss for Basque, Korean & Arabic

We provided the validation loss for masked language modelling on the held-out dev set below for each of the 3 remaining languages ie. Basque (Figure 5), Arabic (Figure 6) and Korean (Figure 7). All the graphs, show a promising trend for the EnerGIZAr set of models, having the lower initial as well as final loss in most comparisons. The FOCUS set of models are often second-best, followed by OFA-768, finally followed by the continual pre-training baseline.

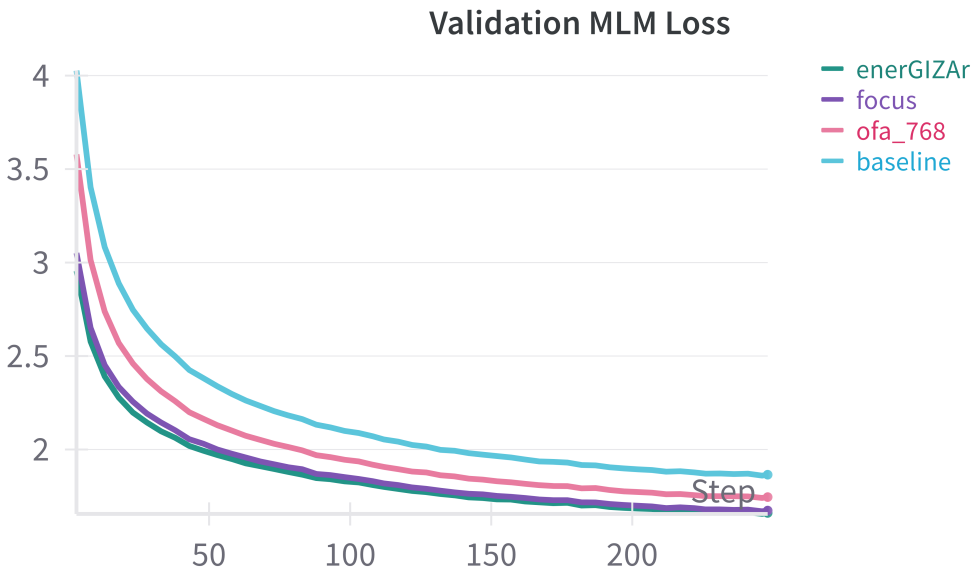

Figure 5: Figure showing the validation masked language modelling loss for Basque wrt. the steps on a held-out dev set for the continual pre-taining baseline, OFA, FOCUS and our EnerGIZAr models.

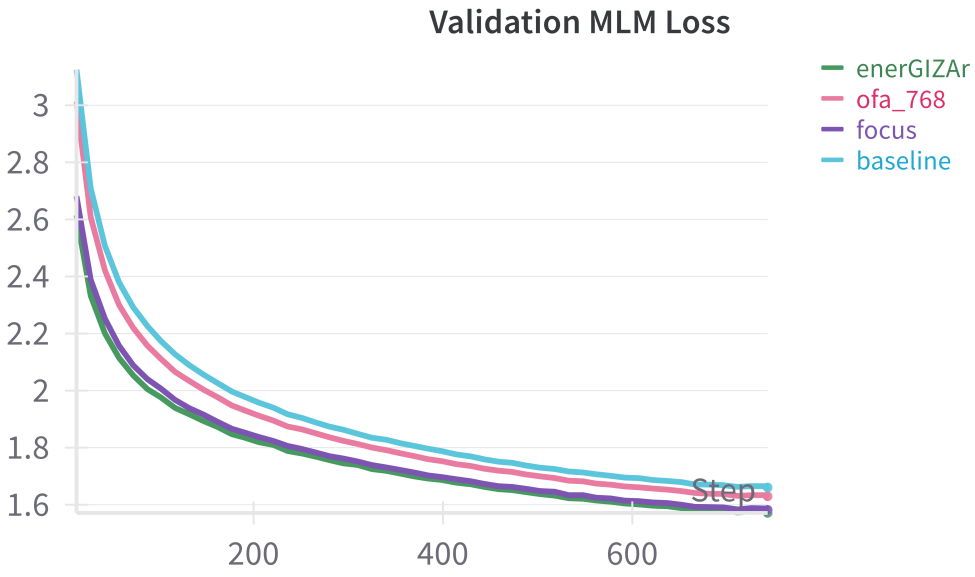

Figure 6: Figure showing the validation masked language modelling loss for Arabic wrt. the steps on a held-out dev set for the continual pre-taining baseline, OFA, FOCUS and our EnerGIZAr models.

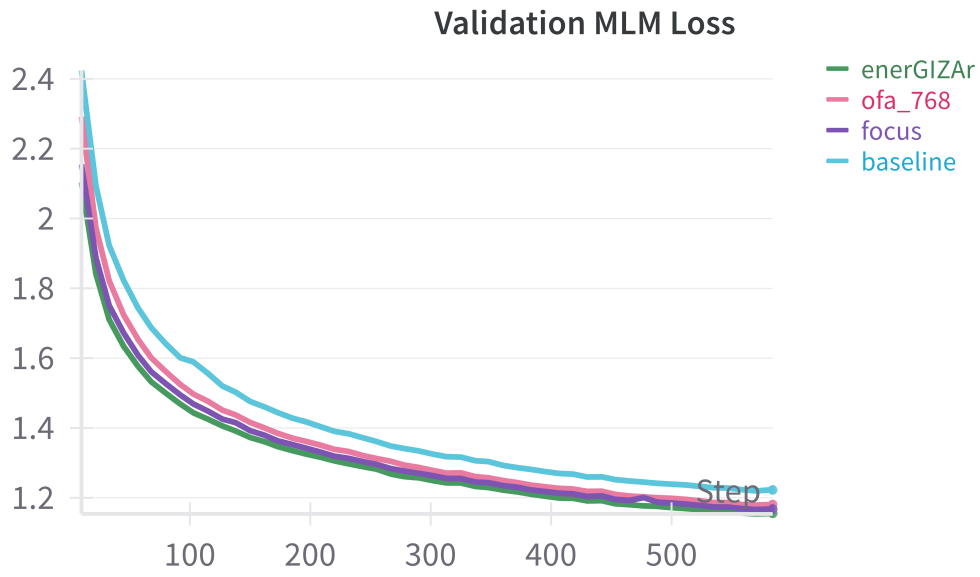

Figure 7: Figure showing the validation masked language modelling loss for Korean wrt. the steps on a held-out dev set for the continual pre-taining baseline, OFA, FOCUS and our EnerGIZAr models.