GreenPLM: Cross-Lingual Transfer of Monolingual Pre-Trained Language Models at Almost No Cost

Qingcheng Zeng[2] , **Lucas Garay**[1] , **Peilin Zhou**[3] , **Dading Chong**[4] , **Yining Hua**[5] , **Jiageng Wu**[1] , **Yikang Pan**[1] , **Han Zhou**[6] , **Rob Voigt**[2] and **Jie Yang**[1]

[1]School of Public Health and the Second Affiliated Hospital, Zhejiang University
[2]Department of Linguistics, Northwestern University
[3]Data Science and Analytics Thrust, Hong Kong University of Science and Technology (Guangzhou)
[4]School of Electronic and Computer Engineering, Peking University
[5]Department of Biomedical Informatics, Harvard University
[6]Language Technology Lab, University of Cambridge
jieynlp@gmail.com

Abstract

Large pre-trained models have revolutionized natural language processing (NLP) research and applications, but high training costs and limited data resources have prevented their benefits from being shared equally amongst speakers of all the world's languages. To address issues of cross-linguistic access to such models and reduce energy consumption for sustainability during large-scale model training, this study proposes an effective and energy-efficient framework called Green-PLM that uses bilingual lexicons to directly "translate" pre-trained language models of one language into another at almost no additional cost. We validate this approach in 18 languages' BERT models and show that this framework is comparable to, if not better than, other heuristics with high training costs. In addition, given lightweight continued pretraining on limited data where available, this framework outperforms the original monolingual language models in six out of seven tested languages with up to 200x less pre-training efforts. Aiming at the Leave No One Behind Principle (LNOB), our approach manages to reduce inequalities between languages and energy consumption greatly. We make our codes and models publicly available at https://github.com/qcznlp/GreenPLMs.

1 Introduction

In recent years, NLP has welcomed great advances and significant progress in deep learning [Vaswani *et al.*, 2017; Devlin *et al.*, 2019; Brown *et al.*, 2020]. Pre-trained language models (PLMs) trained on extensive text have significantly improved performance on various core NLP tasks, bringing NLP new research paradigms such as "pre-training and fine-tuning" and "prompt-based learning" [Liu *et al.*, 2022]. However, despite PLMs' extraordinary performance, training these models both incurs substantial computational costs and relies upon the availability of extensive training data. The former has raised public concern about substantial energy consumption and environmental damage [Strubell *et al.*, 2020], while the latter inevitably creates unequal opportunities for NLP research across languages with differing access to resources [Joshi *et al.*, 2020].

Computation costs for training a single, state-of-the-art deep learning model expanded 300,000 times between 2012 and 2018, a trend which has only intensified since; this presents a challenge for researchers interested in upholding green AI principles [Schwartz *et al.*, 2020] as well as broader public policy ambitions like the United Nations Sustainable Development Goals (SDG) [Assembly, 2015]. Figure 1 shows a comparison of a number of representative PLMs and multimodal models from BERT-base [Peters *et al.*, 2018] to GPT-4 [OpenAI, 2023], where we observe trends of ever-increasing model size, computational costs, economic costs, and environmental costs [Strubell *et al.*, 2019]. Pre-training a BERT-base model incurs cloud computing costs of \$3,751 to \$12,571 [Schwartz *et al.*, 2020], which is unaffordable for personal use and most researchers in developing countries. Training models such as BERT, T5 [Raffel *et al.*, 2020], PaLM, and GPT-3 [Brown *et al.*, 2020] from scratch emit 0.65, 46.7, 271.43, and 552.1 tons of CO_2, respectively.

Beyond economic and environmental costs, the pre-training of large language models requires vast amounts of monolingual text data, which is lacking in many low-resource languages, leading to increased disparities in the utilization of language models among different languages. Languages with a larger number of speakers inherently have larger corpora available for PLM training. This leaves only a tiny fraction of the approximately 7000 languages worldwide with sufficient data to train a monolingual PLM [Ebrahimi and Kann, 2021]. Table 1 lists the sizes of training materials for the monolingual BERT of several languages, exhibiting a significant inequality in monolingual materials. For instance, English GPT-3 was trained on 45 TB of monolingual material, equivalent to 4,608 times size of the Hindi BERT pretraining corpus. This basic inequality in data availability is further exacerbated by many social and financial factors, such
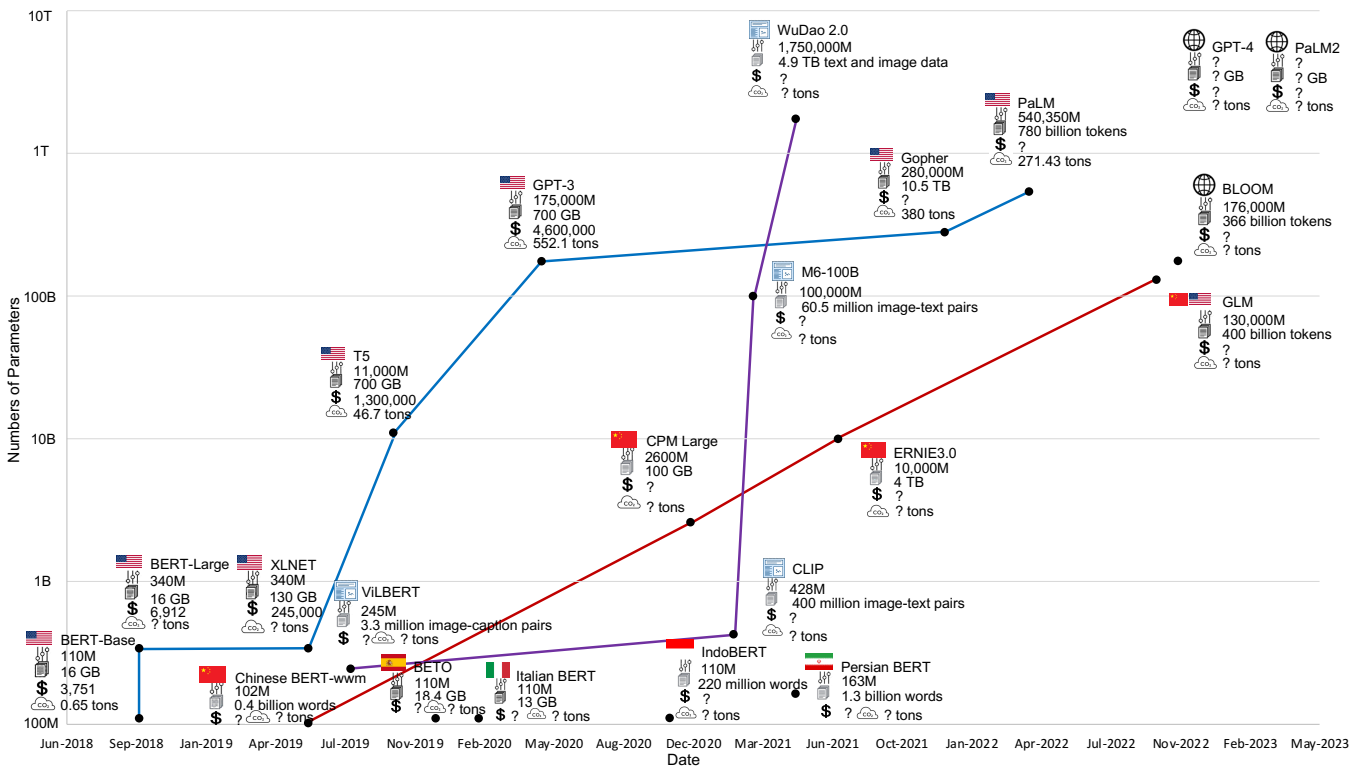
Figure 1: Trends of Complexity and Training Cost in Mainstream PLMs with Significant Performance Boosts. The icons represent the number of parameters, training corpus size, financial cost by dollars, and the emission of CO_2, respectively. ? represents numbers that can not be estimated.

as internet infrastructure, education, and computational resources available in low-resource speech communities. Self-supervised pre-training thus enlarges the disparity between high- and low-resource languages, with the result that breakthrough technologies in contemporary NLP have proven particularly difficult to share equally among speakers of all the world's languages.

To contribute to ongoing efforts to mitigate these issues, we propose a generalizable framework, GreenPLM, to "translate" monolingual PLMs to new languages at almost no additional cost. We hypothesize that the linguistic knowledge learned by PLMs on large monolingual corpora is transferable to low-resource languages, encouraged by two primary pieces of evidence. First, monolingual word embeddings are partially isomorphic across languages [Artetxe *et al.*, 2016]; etymologically close language pairs show good word translation performance [Conneau *et al.*, 2017]. Second, work on the capacities of monolingual PLMs [Conneau *et al.*, 2020b; Blevins and Zettlemoyer, 2022] has shown that they encode cross-linguistically useful information even without training in further languages. Therefore, GreenPLM utilizes bilingual lexicons to bridge cross-lingual semantic spaces while keeping the well-trained parameters of high-resource PLMs fixed as shown in Figure 2. Our experimental results show that GreenPLM's performance is comparable to or better than current heuristics in 18 tested languages, at nearly zero cost. In addition, with around 0.5% computational cost compared

to pre-training a BERT model from scratch, continued pre-training of these models (which we call GreenPLM+) outperforms the performance of monolingual PLMs. To sum up, our contributions to AI and social good are the following:

1) We propose a simple, heuristic pipeline utilizing bilingual lexicons to translate a source language PLM to a target language PLM with almost no computational cost, significantly reducing carbon emissions for building foundation PLMs in various languages;

2) We verify the effectiveness and generalizability of this framework with evaluations in 18 languages, showing performance comparable to monolingual and multilingual PLMs;

3) We propose these models can be further enhanced with continued pre-training in 7 languages and show performance exceeding existing monolingual and multilingual PLMs while reducing pre-training costs up to 200x compared to pre-training from scratch;

4) We suggest this approach can be used to reduce language disparities and energy consumption in pre-training LMs, aligned with the SDG and the LNOB principle to promote the application of AI for social good.
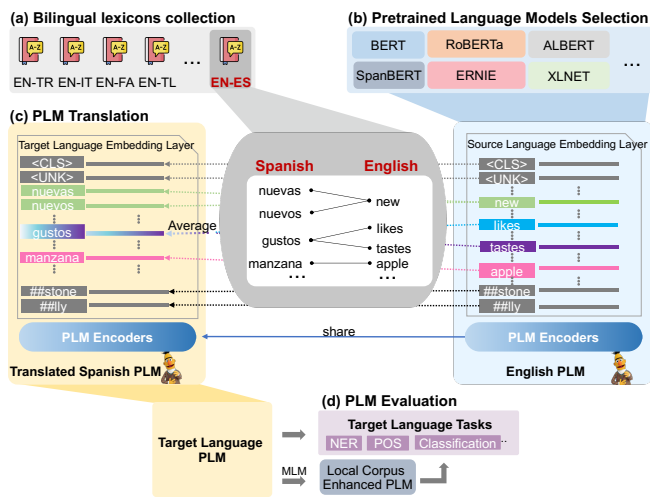
Figure 2: Overview of the GreenPLM approach. Bilingual lexicons (a) are used to augment the vocabulary and embedding layer of a source PLM (b) using our proposed Lexicon Walk Mapping strategy (c), while holding the parameters and architecture of the source model constant, resulting in a translated model that can be further fine-tuned towards particular target-language tasks or continually pre-trained using a smaller dataset in the translated language (d).

2 Related Work

2.1 Cost Reduction in PLM Training

As we've explored, training large PLMs inevitably requires extensive computational resources, which attracted many studies on efficient PLM pre-training. For example, [?] proposed a task-driven language modeling framework to pre-train task-specific PLMs that required only around 2% of the original computational costs. An alternative approach explored by [Chen *et al.*, 2022] and [Qin *et al.*, 2022] leveraged a small "teacher PLM" to pre-train a large "student PLM" at a low computational cost. Reducing data requirements remains less studied compared to the former. The issue of data requirements has been less studied, though recently [Warstadt *et al.*, 2023] put forth a BabyLM challenge to call for sample-efficient pre-training. Our work addresses both these issues, focusing on simultaneously reducing computational costs and data requirements.

2.2 Language Disparity in PLMs

Researchers have proposed numerous mechanisms to equalize performance across human languages in PLMs [Muller *et al.*, 2021; Ebrahimi and Kann, 2021], but the democratization of language technology remains a major challenge [Bird, 2020]. Multilingual BERT (mBERT) represents a preliminary attempt to adapt PLMs to multiple languages and is pre-trained on the concatenation of corpora from different languages [Devlin *et al.*, 2019]. Such models can achieve substantial coverage; for example, mBERT and XLM-R [Conneau *et al.*, 2020a] cover 104 and 100 languages respectively. Nevertheless, model performances in specific languages are reported to be highly dependent on pre-training corpus sizes, and cross-lingual PLMs trained in this manner are still expensive to train and ultimately rely upon the availability of monolingual texts in a given target language.

2.3 Cross-Lingual Transfer

After multilingual BERT showed impressive performance in zero-shot cross-language transfer settings [Pires *et al.*, 2019], substantial attempts have been made to adapt monolingual or multilingual PLMs to other languages.

[Wu and Dredze, 2020] examined the performance of mBERT across various languages and reported that the model performs well for high-resource languages but not for low-resource languages. Nonetheless, in low-resource settings, mBERT performs better than monolingual BERT, validating some degree of cross-language generalizability. [Muller *et al.*, 2021] explored mBERT's performance in never-before-seen languages, finding unstable performance. They found that mBERT has unstable performance across languages, depending on the training data. [Ebrahimi and Kann, 2021] used the New Testament, which is available in most languages, to adapt multilingual PLMs to more than 1600 languages using continued pre-training, vocabulary extension, and adapter transformers. [Wang *et al.*, 2022] explored how bilingual lexicons can enhance the multilingual language coverage of PLMs. Their results had consistent improvement across sequence labeling and dependency parsing tasks in 19 low-resource languages. In contrast, our current study extends bilingual lexicons to monolingual settings as a bridge rather than data augmentation components to design specified monolingual PLMs equally for all languages.

Monolingual adaptation remains relatively under-explored compared to multilingual. It follows two general directions: extending monolingual PLMs to bilingual and adapting a source language PLM to a target language PLM. In the first direction, [Artetxe *et al.*, 2020] continually pre-trained English PLMs on data in target languages with the transformer body frozen and fine-tuned the PLMs on labeled English data. Their results confirmed the generalizability of monolingual PLMs across languages. [Tran, 2020] transferred PLMs by initializing foreign language embeddings in the English vector space and jointly fine-tuning both models. For the second direction, [de Vries and Nissim, 2021], and [Minixhofer *et al.*, 2022] used word embeddings as the bridge to adapt monolingual models to other languages with continued pre-training. Motivated by these works, we leverage more widely available and less computationally expensive resources other than word embeddings to generalize PLMs.

3 Methodology

3.1 Overall Architecture

An overview of the GreenPLM framework is shown in Figure 2. We start from a bilingual lexicon for a given source-target language pair and the selection of a source PLM. Importantly, contemporary PLMs require a pre-established vocabulary embedding layer which operates on the outputs of a specific tokenizer. Tokens in the vocabulary are a mix of both full words and word pieces, traditionally the outputs of an algorithm like WordPiece or byte pair encoding, where each token is mapped to an embedding learned in the pre-training process.

In our approach, we augment the vocabulary and embedding layer of the source PLM to incorporate target-language vocabulary from the bilingual lexicon, using a method to map this lexicon into a shared semantic space explained below in subsection 3.2. This results in what we call a "translated PLM," though notably the embedding layer still contains the entire vocabulary from the source language and the PLM encoder parameters are yet unchanged from the source PLM. Such a model can already be applied directly to target-language data; in experiments, we call this setting simply "GreenPLM".

The translated PLM can also be fine-tuned towards particular tasks in the target language, or pre-trained further on target language data, both of which modify the PLM parameters to be more suited to the characteristics of the target language. We name this setting "GreenPLM+".

3.2 Lexicon Walk Mapping

The biggest challenge in the approach outlined above is how to use a bilingual lexicon to generate coherent target-language embeddings that can be fed into source-language PLM encoder parameters. To achieve this we propose a heuristic method we call Lexicon Walk Mapping (LWM), in which a bilingual lexicon is used as the bridge to create a mapping between the semantic spaces of the source and target languages.

We define **a bilingual lexicon** as $L = \{x_i, y_i\}$ where x_i is a word or phrase in the source language (English, in this work) and y_i is x_i's translation in the target language. Then given a **source language's PLM S** with its vocabulary S_{vocab} and token embeddings $S_{embedding}$, we walk through the lexicon and operate on each entry separately. For each entry, we first tokenize the source word or phrase x_i into a set of tokens $T = \{t_1, t_2, ...t_i...t_n \mid t_i \in S_{vocab}\}$, and then map T to the token embedding layer of S to retrieve a set of embeddings $E = \{e_1, e_2, ...e_i...e_n \mid e_i \in S_{embedding}\}$. We then take the dimension-wise average of E, and the new embedding $e_{average}$ serves as y_i's token embedding in the Green-PLM. If y_i repeats in the lexicon, as is frequently the case in one-to-many translations across languages, the GreenPLM framework accumulates several instances of $e_{average}$ and in turn averages them at the end of the procedure to generate a final token embedding.

Our approach leaves the original token embeddings of source language words, tokenized word pieces, and special characters in S_{vocab} untouched in the translated PLM; instead, we merely expand the vocabulary size by adding unseen target language words in the bilingual lexicon. For example in Spanish, the GreenPLM has a vocabulary size of 119,999 words, while the corresponding Spanish monolingual BERT only has 31,002.

3.3 Continued Pre-training

Continued pre-training has long been validated as one of the most effective strategies to enhance PLMs, with several possible strategies including masked language modeling (MLM) [Devlin et al., 2019] and translation language modeling (TLM) [Conneau and Lample, 2019]. In addition, some special techniques, such as contrastive learning, can be combined with pre-training tasks. Here we use MLM as the basic pre-training task in our continued pre-training experiments. To generate a GreenPLM+ model for each language, the translated PLM is further pre-trained with a batch size of 1024 and a sequence length of 128 symbols by utilizing 10% of the OSCAR dataset [Ortiz Suárez et al., 2019]. Each model was pre-trained for 5 epochs and every 10,000 pre-training steps were saved.

The experiments presented here examine GreenPLM+ models in seven languages: Spanish, Indonesian, Tagalog, Turkish, Italian, Romanian, and Persian because they are the only languages with a moderate amount of language resources. An attempt to further pre-train a GreenPLM for Maltese fails because the available resources (17 MB in the OSCAR corpus) are too scarce to maintain training stability.

3.4 Alternatives to LWM

We found that LWM worked most effectively as a method to bridge the embedding spaces between models, and therefore present experimental results using that method. However, we also tried several other logical alternatives that were not as effective. We describe those methods here for completeness and report detailed results of these methods in Spanish, Turkish, and Tagalog in the supplementary material. Because these methods are strictly worse, we did not further experiment on other languages.

Vocabulary Expansion (VE) refers to expanding the vocabulary based on the vocabulary of the source language PLM. Given a **bilingual lexicon** $lexicon = \{x_i, y_i\}$ and the PLM S of the source language with a vocabulary of $S_{vocab} = \{w_1, w_2, ..., w_i, ..., w_n\}$ and token embeddings $S_{embedding}$, VE will walk through S_{vocab} and find their corresponding embeddings by $S_{embedding}$ and $lexicon$. If no matching entries exist, this strategy will keep the original tokens. If there is one y_i corresponding to multiple x_i, VE will keep the first x_i's embedding as the translated embedding. Then, VE will merge all W_i as the vocabulary of the new model, where translated embeddings are used for the newly extended vocabulary.

Vocabulary One-on-one Mapping (VOM) adopts a similar methodology to VE but only uses one translation for multiple entries. While both LWM and VE increase the vocabulary sizes, introducing more parameters, VOM yields a Green-PLM with the same vocabulary sizes.

Vocabulary Translation Mapping (VTM) is the most basic strategy being tested. It cuts sentences according to space delimitation and uses bilingual lexicons to directly transfer the tokens into another language. Then, the newly generated sentences are directly fed to the source language PLM for downstream tasks.

4 Materials and Evaluation

4.1 Language and Model Choices

English was chosen as a source language since it has the largest pre-training corpus size and the largest body of existing PLMs available. For target languages, we included 18 languages from five language families (Table 1). According to [Joshi et al., 2020], the chosen languages fall under various language resources' levels ranging from the extremely low-

Language	Language Family	Language Resource	Monolingual PLMs' Corpus Size
Spanish	Indo-European	5 - Winners	2996.01 million words
Indonesian	Austronesian	3 - Rising Stars	220 million words
Turkish	Turkic	4 - Underdogs	440.50 million words
Tagalog	Austronesian	3 - Rising Stars	39 million words
Italian	Indo-European	4 - Underdogs	2050 million words
Romanian	Indo-European	3 - Rising Stars	2421.3 million words
Persian	Indo-European	4 - Underdogs	1300 million words
Maltese	Afro-Asiatic	2 - Hopefuls	-
Māori	Austronesian	1 - Scraping-Bys	-
Wolof	Niger–Congo	2 - Hopefuls	-
Luganda	Niger–Congo	1 - Scraping-Bys	-
Ilokano	Austronesian	1 - Scraping-Bys	-
Hausa	Afro-Asiatic	2 - Hopefuls	-
Bulgarian	Indo-European	3 - Rising Stars	-
Latvian	Indo-European	3 - Rising Stars	-
Ancient Greek	Indo-European	-	-
Norwegian	Indo-European	1 - Scraping-Bys	-
Danish	Indo-European	3 - Rising Stars	-

Table 1: Languages represented in this work, representing five language families and a variety of settings with regard to resource availability.

resource level "1 - The Scraping-Bys" to the high-resource level "5 - The Winners".

In these experiments, we used the *bert-base-uncased* English model as the source PLM, since it has the broadest coverage in academic and industrial usage. Since the model is relatively lightweight it also allows for extensive experiments on various NLP tasks across languages, even with limited computational resources. Nevertheless, the GreenPLM framework is independent of the particular structure of PLMs beyond the embedding layer, and hence can theoretically be applied to larger models to benefit from continuing advances in large-scale PLMs.

4.2 Experimental Settings

Bilingual Lexicons

This study uses bilingual lexicons from four sources. For relatively high-resource languages, our lexicons are from MUSE [Conneau *et al.*, 2017], PanLex [Kamholz *et al.*, 2014], and OPUS [Tiedemann and Nygaard, 2004]. For Maltese, Māori, Wolof, Luganda, Ilokano, and Hausa, we used lexicons released from [?] directly. Detailed statistics of the lexicons could be accessed in the supplementary material.

Evaluation

Most chosen languages have available benchmark datasets for performance comparison. For example, [Cañete *et al.*, 2020] as the Spanish benchmark; IndoLEM in [Koto *et al.*, 2020] for Indonesian; [Schweter, 2020a], [Cruz and Cheng, 2019; Cruz and Cheng, 2020; Cruz *et al.*, 2020], [Schweter, 2020b], [Dumitrescu *et al.*, 2020] and [Farahani *et al.*, 2021] for Turkish, Tagalog, Italian, Romanian, and Persian, respectively. For the 11 low-resource languages with only NER and POS tasks, GreenPLM was tested on MasakhaNER [Adelani *et al.*, 2021] and universal dependencies [de Marneffe *et al.*, 2021]. For benchmark datasets with multiple tasks, we divided the tasks into three main categories: POS, NER, and text classification for a by-category analysis of our GreenPLM framework. Specifically, for continued pre-training, this study uses *batch size × sequence length × training steps* as a metric to evaluate pre-training efforts.

Our translated PLMs were evaluated by performance comparison with existing models on various NLP tasks, includ-

ing part-of-speech tagging (POS), named entity recognition (NER), sentiment analysis, sentence classification, and paraphrase identification. We used the F1 score for the NER task, accuracy in POS and most classification tasks, Pearson's coefficient in the tweet ordering task of Indonesian, and hamming loss in the Dengue task of Tagalog. Specifically, We re-scaled rare metrics like hamming loss and Pearson's coefficient to fit a 0 to 100 scale and took the average of model performance for a direct comparison.

Baseline Methods

To compare the performance of GreenPLM with existing models, four deep learning models were chosen as baselines, and fine-tuned towards each evaluation task: (1) Scratch: a randomly initialized BERT model; (2) Pre-BERT SOTA: state-of-the-art models before BERT, i.e. long short-term memory (LSTM) and convolutional neural networks (CNNs) with pre-trained word embeddings, which we implement using NCRF++ [Yang and Zhang, 2018]; (3) mBERT: multilingual BERT [Devlin *et al.*, 2019]; and (4) monolingual BERT models, where available. In addition, GreenPLM+, the advanced version of GreenPLM with a continued pre-training on a very limited dataset, was also compared with the above models. Only uncased models were tested in this study because most bilingual lexicons are uncased.

5 Results

Our experiments test baseline, GreenPLM, and GreenPLM+ models on 18 languages across various categories of tasks listed in Table 2. Detailed results on each specific task could be accessed in the *supplementary material*.

5.1 Base GreenPLM

In the base setting where models are translated using only LWM, we observe that the GreenPLM performs much better than random BERT in all cases, and better than Pre-BERT SOTA models in 15 out of 18 languages, with the exception of sequence labeling tasks in 3 languages. This suggests that indeed, the bilingual lexicon based approach we take with GreenPLM is not merely successful due to task-specific fine-tuning, but rather achieves performance gains due to cross-linguistic knowledge transfer from the source PLM. Though GreenPLM generally does not surpass mBERT in performance, our method provides at least comparable performance on the examined languages at nearly no cost, whereas mBERT itself requires large multilingual training costs.

Moreover, in some language settings our method is surprisingly effective. In **Tagalog**, the GreenPLM wins over all baselines and shows a large improvement over the monolingual BERT. This may imply that the original Tagalog BERT was not pre-trained with sufficient data. In the low-resource settings of **Maltese** and **Ilokano**, the GreenPLM also outperforms all existing baselines.

5.2 GreenPLM+: Continued Pre-training

As a further enhancement to the GreenPLM, we applied a lightweight continued pre-training step with very limited

Language	Task by Category	Scratch	Pre-BERT SOTA	mBERT	monolingual BERT	GreenPLM	GreenPLM+
Spanish	POS, NER, Classification*3	67.38	81.68	88.04	**89.37**	85.27	88.08 (0.7%)
Indonesian	POS, NER*2, Classification*3	63.67	75.17	81.09	84.56	81.08	**84.57** (1.4%)
Turkish	POS*2, NER	70.38	90.41	92.97	93.07	90.70	**93.09** (2.5%)
Tagalog	Classification*3	77.41	77.44	89.43	85.77	89.59	**89.59** (0%)
Italian	POS*2	83.47	96.39	97.44	97.38	95.98	**97.63** (0.5%)
Romanian	POS, NER, Classification*2	67.08	82.26	86.5	88.72	85.39	**88.79** (1.2%)
Persian	NER*2, Classification*3	84.9	90.52	91.96	93.59	91.05	**93.60** (0.8%)
Maltese	POS, NER	40.47	70.79	73.30	-	**76.93**	-
Māori	NER	53.33	**89.44**	81.97	-	79.21	-
Wolof	POS, NER	48.45	**72.99**	72.72	-	72.71	-
Luganda	NER	30.39	**74.65**	74.19	-	74.48	-
Ilokano	NER	29.65	64.58	67.44	-	**67.98**	-
Hausa	NER	50.66	82.38	**85.58**	-	84.97	-
Bulgarian	POS	92.16	97.93	**99.06**	-	97.96	-
Latvian	POS	89.53	94.48	**96.33**	-	94.59	-
Ancient Greek	POS	89.42	95.00	**96.00**	-	95.08	-
Norwegian	POS	92.82	96.50	**97.61**	-	96.75	-
Danish	POS	82.37	94.50	**97.80**	-	95.90	-

Table 2: Tasks and experimental results in six settings. Numbers represent average performance across language-specific evaluation tasks; since the available tasks are different for each language, numbers are only comparable within their own row. Parentheticals in the last column present the continued pre-training efforts required to reach performance equal to the corresponding monolingual model compared to training from scratch; for Tagalog, this number is 0% because we didn't continue pretraining GreenPLM as it has outperformed the monolingual one.

training cost and evaluated the results in the seven languages mentioned above with access to sufficiently-sized pre-training corpora. In the last column of Table 2, we evaluate GreenPLM+ models at the first point at which they outperform mBERT or monolingual BERT for the first time during the continued pre-training, and note in parens the percentage of pre-training efforts required to reach this level of performance compared to pre-training a PLM from scratch. We observe that GreenPLM+ outperforms all monolingual BERT at a relatively low cost, except in Spanish. For example, Green-PLM+ outperforms the original PLM with only **0.68%** of the pre-training efforts for Indonesian. For Spanish, Green-PLM+ outperforms multilingual BERT and achieves comparable performance over the monolingual BERT when given **0.7%** of the pre-training efforts. We further compared continuously pre-training GreenPLM and mBERT under the same settings in Indonesian, Romanian, and Persian. GreenPLM could catch up and outperform mBERT given minimal costs. In Figure 3(b), we show a visualized comparison of Persian.

6 Discussion

Impact on Sustainable Development. High computational requirements for pre-training LMs or even pre-BERT deep learning models have been a major obstacle for researchers, where a slight performance boost may come at the expense of dramatically increased model complexity. Our approach generates translated PLMs for medium-level and low-level resource languages that successfully outperform pre-BERT SOTA deep learning models at extremely low pre-training costs, both in terms of economic costs and CO_2 emissions. In contrast to prior approaches for low-cost training of PLMs [Yao *et al.*, 2022], the GreenPLM framework is task-agnostic All computation for PLM translation in the basic GreenPLM approach can be executed on a personal computer within 10 seconds. For continued pre-training with the GreenPLM+

model, we demonstrate comparable results to a fully pre-trained monolingual model using less than 10% of the dataset size in less than 8 hours on an 8x RTX3090 system.

Impacts on Language Equality. Unequal language resources have limited recent advances in PLMs from being equally shared by speakers of all the world's languages. While developing larger corpora for pre-training is not always feasible, here we leverage a more widely available language resource – bilingual lexicons – to translate PLMs from higher- to lower-resource settings. This framework can serve as a basic starting point to transfer models across languages and reduce both language-based and economic disparities in the availability of PLMs in different linguistic and socioeconomic backgrounds.

Experimental results show that GreenPLM+ performs better than existing monolingual and multilingual models in languages with small but still substantial pre-training corpus sizes, such as Indonesian, Turkish, and Tagalog. These languages are rated as "4 - Underdogs" or "3 - Rising Stars" in the [Joshi *et al.*, 2020] classification of resource availability. We note with an example from Persian in Figure 3 (b) that in these settings, the base GreenPLM model commonly starts off with lower performance than mBERT, but the GreenPLM+ quickly catches up after continued pre-training with the same number of steps.

For languages with fewer resources in the '2 - Hopefuls' and '1 - Scraping-Bys' categories by [Joshi *et al.*, 2020], where substantial pre-training corpora are not available for continued pre-training, we find that mBERT retains better performance than the base GreenPLM. Nevertheless, the simple and nearly zero-cost transfer method presented here achieves performance generally surpassing Pre-BERT SOTA models, and comparable to mBERT, suggesting that base GreenPLM models can be productively used to transfer PLMs to languages where only bilingual lexicons exist.
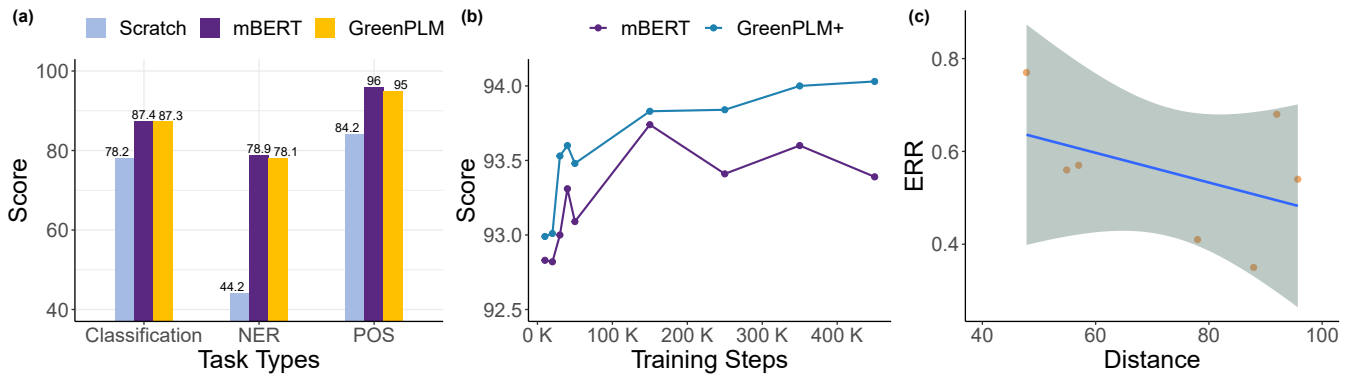
Figure 3: Performance comparison of GreenPLM under different settings. **(a)** Results across NER, POS, and Classification task types. **(b)** Comparison between continued pre-training GreenPLM+ versus multilingual BERT in Persian. **(c)** We observe a linear relationship between language distance and error reduction rate.

Applicability for Downstream Tasks. The evaluations in this work examined downstream tasks in the categories of POS tagging, NER, and classification tasks, corresponding to word-level, span-level, and sentence-level understanding. Results in Figure 3(a) show that the GreenPLM, compared to random BERT models, has the most remarkable improvement on POS tasks, followed by NER and classification tasks. However, compared to mBERT, GreenPLM falls short in POS, followed by NER and classification tasks. This suggests that GreenPLM transfers sentence-level understanding the best in downstream tasks, possibly because embedding-based strategies of our framework focus mainly on semantic information and lack the ability to capture the variability of syntactic information in natural languages.

Language Distance. In addition, we use a language distance calculator from eLinguistics [1] as a reference and quantify the performance by the error reduction rate (ERR) metric because languages are classified according to their diachronic relatedness in linguistic typology. Corresponding results in Figure 3(c) show that the closer the selected language is to English, the better its GreenPLM will perform. This is not surprising since GreenPLM models retain the parameters from the original English BERT being translated; nevertheless, it suggests that linguistic distance is likely to be a crucial consideration in cross-lingual knowledge transfer, supporting existing findings [Lin *et al.*, 2019].

Future Work. Though this paper uses English as a source language and starts from an English PLM to generate target language PLMs, a similar approach could be used to merge PLMs from multiple languages for cross-enhancement to generate a single target language PLM. In addition, GreenPLM could be extended to multi-modal and multi-domain settings where bilingual dictionaries could facilitate information exchange; for example, could a similar approach help integrate biomedical information using a targeted lexicon? Finally, it remains to investigate how the GreenPLM approach performs with more state-of-the-art BERT-class PLMs such as RoBERTa, as well as with PLMs of different architectures

such as BART [Lewis *et al.*, 2020].

Limitations. This work has two main limitations. First, it only covers uncased PLMs at the current stage because most bilingual lexicons are uncased, and capitalization could affect model performance. Second, there is no mechanism to rank the importance of single entries in bilingual lexicons. Therefore, newly generated models generally have a more extensive vocabulary than the original PLM.

7 Conclusion

This paper proposes a bilingual lexicon-based framework GreenPLM that "translates" a monolingual PLM to other languages with no additional cost. We verify its effectiveness in 18 languages and further pre-train the resulting models to validate their performance when given minimal pre-training costs. Our results provide evidence that the GreenPLM+ (GreenPLM with very limited continued pre-training) outperforms monolingual BERT at much lower computational costs than current frameworks, such as the continued pre-training of mBERT. This framework is easy to incorporate into existing PLMs and can be adapted to various languages given the relatively wide availability of bilingual lexicons. We hope that this work can make a positive impact on promoting green AI and reducing language disparities in access to contemporary NLP technologies.

Acknowledgments

Contribution Statement

Qingcheng Zeng, Lucas Garay, and Peilin Zhou contributed equally to this paper. Correspondence to Jie Yang.

References

[Adelani *et al.*, 2021] David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe

---

[1] http://www.elinguistics.net/Compare_Languages.aspx

Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021.

[Artetxe *et al.*, 2016] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 EMNLP*, pages 2289–2294, Austin, Texas, November 2016.

[Artetxe *et al.*, 2020] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th ACL*, pages 4623–4637, Online, July 2020.

[Assembly, 2015] UN General Assembly. Transforming our world: the 2030 agenda for sustainable development. https://www.refworld.org/docid/57b6e3e44.html, 2015. Accessed: 2023-06-13.

[Bird, 2020] Steven Bird. Decolonising speech and language technology. In *Proceedings of the 28th COLING*, pages 3504–3519, Barcelona, Spain (Online), December 2020.

[Blevins and Zettlemoyer, 2022] Terra Blevins and Luke Zettlemoyer. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 EMNLP*, pages 3563–3574, Abu Dhabi, United Arab Emirates, December 2022.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *NIPS*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[Cañete *et al.*, 2020] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pretrained bert model and evaluation data. In *PML4DC at ICLR 2020*, 2020.

[Chen *et al.*, 2022] Cheng Chen, Yichun Yin, Lifeng Shang, Xin Jiang, Yujia Qin, Fengyu Wang, Zhi Wang, Xiao Chen, Zhiyuan Liu, and Qun Liu. bert2BERT: Towards reusable pretrained language models. In *Proceedings of the 60th ACL*, pages 2134–2148, Dublin, Ireland, May 2022.

[Conneau and Lample, 2019] Alexis Conneau and Guillaume Lample. *Cross-Lingual Language Model Pretraining*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[Conneau *et al.*, 2017] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

[Conneau *et al.*, 2020a] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th ACL*, pages 8440–8451, Online, July 2020.

[Conneau *et al.*, 2020b] Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th ACL*, pages 6022–6034, Online, July 2020.

[Cruz and Cheng, 2019] Jan Christian Blaise Cruz and Charibeth Cheng. Evaluating language model finetuning techniques for low-resource languages. *arXiv preprint arXiv:1907.00409*, 2019.

[Cruz and Cheng, 2020] Jan Christian Blaise Cruz and Charibeth Cheng. Establishing baselines for text classification in low-resource languages. *arXiv preprint arXiv:2005.02068*, 2020.

[Cruz *et al.*, 2020] Jan Christian Blaise Cruz, Jose Kristian Resabal, James Lin, Dan John Velasco, and Charibeth Cheng. Investigating the true performance of transformers in low-resource languages: A case study in automatic corpus creation. *arXiv preprint arXiv:2010.11574*, 2020.

[de Marneffe *et al.*, 2021] Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, 07 2021.

[de Vries and Nissim, 2021] Wietse de Vries and Malvina Nissim. As good as new. how to successfully recycle English GPT-2 to make models for other languages. In *Findings of the ACL: ACL-IJCNLP 2021*, pages 836–846, Online, August 2021.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 NAACL-HLT*, pages 4171–4186, Minneapolis, Minnesota, June 2019.

[Dumitrescu *et al.*, 2020] Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. The birth of Romanian BERT. In *Findings of the ACL: EMNLP 2020*, pages 4324–4328, Online, November 2020.

[Ebrahimi and Kann, 2021] Abteen Ebrahimi and Katharina Kann. How to adapt your pretrained multilingual model to 1600 languages. In *Proceedings of the 59th ACL*, pages 4555–4567, Online, August 2021.

[Farahani *et al.*, 2021] Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. ParsBERT: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53(6):3831–3847, oct 2021.

[Joshi *et al.*, 2020] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th ACL*, pages 6282–6293, Online, July 2020.

[Kamholz *et al.*, 2014] David Kamholz, Jonathan Pool, and Susan Colowick. PanLex: Building a resource for panlingual lexical

translation. In *Proceedings of the Ninth LREC*, pages 3145–3150, Reykjavik, Iceland, May 2014.

[Koto *et al.*, 2020] Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th COLING*, pages 757–770, Barcelona, Spain (Online), December 2020.

[Lewis *et al.*, 2020] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th ACL*, pages 7871–7880, Online, July 2020.

[Lin *et al.*, 2019] Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th ACL*, pages 3125–3135, Florence, Italy, July 2019.

[Liu *et al.*, 2022] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, sep 2022.

[Minixhofer *et al.*, 2022] Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 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, July 2022. Association for Computational Linguistics.

[Muller *et al.*, 2021] Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 NAACL-HLT*, pages 448–462, Online, June 2021.

[OpenAI, 2023] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[Ortiz Suárez *et al.*, 2019] Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 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, 2019. Leibniz-Institut für Deutsche Sprache.

[Peters *et al.*, 2018] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 NAACL-HLT*, pages 2227–2237, New Orleans, Louisiana, June 2018.

[Pires *et al.*, 2019] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th ACL*, pages 4996–5001, Florence, Italy, July 2019.

[Qin *et al.*, 2022] Yujia Qin, Yankai Lin, Jing Yi, Jiajie Zhang, Xu Han, Zhengyan Zhang, Yusheng Su, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Knowledge inheritance for pre-trained language models. In *Proceedings of the 2022 NAACL-HLT*, pages 3921–3937, Seattle, United States, July 2022.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou,

Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67, 2020.

[Schwartz *et al.*, 2020] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63(12):54–63, 2020.

[Schweter, 2020a] Stefan Schweter. Berturk - bert models for turkish. https://doi.org/10.5281/zenodo.3770924, 2020. Accessed: 2023-06-13.

[Schweter, 2020b] Stefan Schweter. Italian bert and electra models. https://doi.org/10.5281/zenodo.4263142, 2020. Accessed: 2023-06-13.

[Strubell *et al.*, 2019] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th ACL*, pages 3645–3650, Florence, Italy, July 2019.

[Strubell *et al.*, 2020] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. *Proceedings of the AAAI*, 34(09):13693–13696, Apr. 2020.

[Tiedemann and Nygaard, 2004] Jörg Tiedemann and Lars Nygaard. The OPUS corpus - parallel and free: http://logos.uio.no/opus. In *Proceedings of the Fourth LREC*, Lisbon, Portugal, May 2004.

[Tran, 2020] Ke Tran. From english to foreign languages: Transferring pre-trained language models. *arXiv preprint arXiv:2002.07306*, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[Wang *et al.*, 2022] Xinyi Wang, Sebastian Ruder, and Graham Neubig. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[Warstadt *et al.*, 2023] Alex Warstadt, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox, and Chengxu Zhuang. Call for papers–the babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2301.11796*, 2023.

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

[Yang and Zhang, 2018] Jie Yang and Yue Zhang. NCRF++: An open-source neural sequence labeling toolkit. In *Proceedings of ACL 2018, System Demonstrations*, pages 74–79, Melbourne, Australia, July 2018.

[Yao *et al.*, 2022] Xingcheng Yao, Yanan Zheng, Xiaocong Yang, and Zhilin Yang. NLP from scratch without large-scale pretraining: A simple and efficient framework. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th ICML*, volume 162 of *PMLR*, pages 25438–25451. PMLR, 17–23 Jul 2022.