A Balanced Data Approach for Evaluating Cross-Lingual Transfer: Mapping the Linguistic Blood Bank

Anonymous ACL submission

Abstract

We show that the choice of pretraining languages affects downstream cross-lingual transfer for BERT based models. We inspect zero-shot performance under balanced data conditions to mitigate data size confounds, classifying pretrain languages that increase downstream performance as *donors*, and languages that are most improved in zero-shot performance as *recipients*. We develop a method of quadratic time complexity in the number of pretraining languages to estimate these inter-language relations, instead of an exponential exhaustive computation of all possible combinations. We find that our method is effective on a diverse set of languages spanning different linguistic features and two downstream tasks. Our findings can inform developers of future large-scale multilingual language models in choosing better pretraining configurations.

1 Introduction

Pretrained language models (PLMs; Peters et al., 2018; Devlin et al., 2019, inter alia) have been setting state-of-the-art results in most NLP tasks, thanks to their ability to leverage naturally occurring raw texts during pretraining. Interestingly, various recent works found that when pretraining on corpora composed of different languages, PLMs exhibit *zero-shot cross-lingual abilities*, i.e., they achieve non-trivial performance when tested on downstream examples for languages seen only during pretraining. For example, in Figure 1, a part-of-speech model finetuned only on English is capable of predicting correctly on an example in Russian, which was seen only during pretraining (Wu et al., 2019; Wang et al., 2019; Conneau et al., 2019; Lazar et al., 2021; Turc et al., 2021).

Previous analyses have examined how several factors contribute to this emerging behavior. For example, parameter sharing and model depth seem important in certain configurations (Wang et al., 2019; Wu et al., 2019), and the choice of finetune language can lead to varying results (Turc et al., 2021).

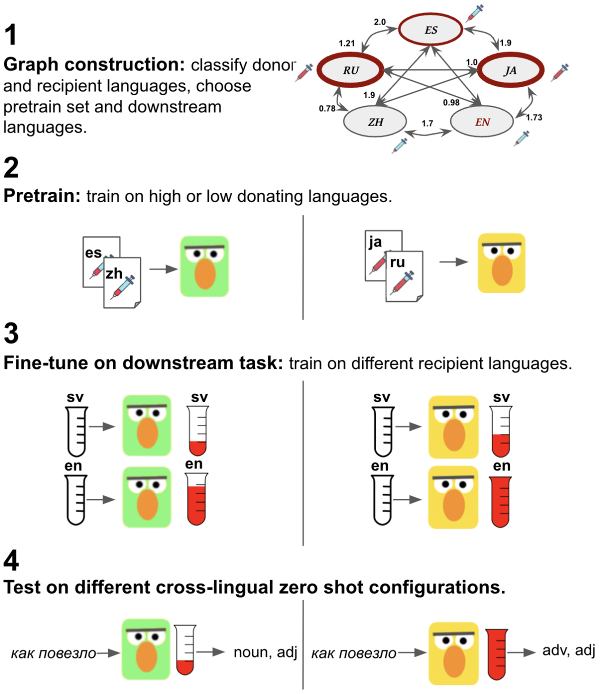

1
Graph construction: classify donor and recipient languages, choose pretrain set and downstream languages.
2.0
ES
1.21
1.9
RU
1.0
JA
1.9
0.78
0.98
1.73
ZH
1.7
EN
2
Pretrain: train on high or low donating languages.
es
zh
ja
ru
3
Fine-tune on downstream task: train on different recipient languages.
sv
sv
en
en
sv
sv
en
en
4
Test on different cross-lingual zero shot configurations.
как повезло
noun, adj
как повезло
adv, adj

Figure 1: Previous studies on cross-lingual transfer focus on choosing transfer languages for fine-tuning while fixing the pretraining phase to e.g., to mBERT's language and data selection. We identify that pretraining plays a major role in cross-lingual performance. In stage 1, we construct a pretraining-based language interaction graph to identify good pretraining and downstream languages (termed "donors" and "recipients") , Following, in stages 2 and 3 we test various combinations of pretraining and downstream configurations, and in stage 4 we show that donor languages indeed lead to better zero-shot performance (indicated by the full beaker) in recipient languages in two downstream tasks.

In this work, we focus on an important factor which we find missing in prior work, namely the effect that *pretraining* languages have on downstream zero-shot performance. In particular, we ask three major research questions: (1) Does the choice of pretraining languages affect downstream cross-lingual transfer, and if so, to what extent? (2) Is English the optimal pretraining language, when controlling for confounding factors such as data size and domain? And finally, (3) Can we choose pretraining languages to improve downstream zero-shot performance?

These questions have implications for the next generation of PLMs, which can find better and more efficient choices for pretraining languages; for downstream model developers in low-resource setting, which can find specific pretraining configurations best suited for their setting; and can help shed light on the inter-relations between languages and how they manifest in cross-lingual transfer.

To answer these research questions, we decouple as much as possible the *language* from its corresponding *dataset*. To the best of our knowledge, prior work has conflated pretrain corpus size and its domain with other examined factors, thus skewing results towards over-represented languages, such as English or German (Joshi et al., 2020).[1]

To achieve this, we first construct a *linguistically-balanced* pretraining corpus based on Wikipedia, composed of a diverse set of 22 languages. We carefully control for the amount of data and domain distribution in each of the languages (Section 3).

Next, we note that the number of pre-training configurations grows exponentially with the number of languages n represented in the dataset.[2] Therefore, it is infeasible to exhaustively test all possible configuration, much less extend it for more languages.

To overcome this limitation, in Section 4 we propose a novel pretraining-based approach which is quadratic in the number of languages and yields an estimation of how much a language is projected to contribute to zero shot performance, based only language modelling performance. Using this approach, we identify four sets of languages. Two sets of *pretrain* languages: (1) *top-donating* – languages which we expect would yield the best downstream zero-shot performance, (2) *least-donating* – languages which we expect would lead to relatively worst zero-shot performance, and two sets of *downstream* languages: (3) *most recipient*, languages which we expect would show best zero-shot performance, and (4) *least recipient*, languages which we expect would lead to worst zero-shot transfer.

Finally, our evaluations on two multilingual downstream tasks (part of speech tagging and named entity recognition) lead to three main conclusions (Section 5): (1) the choice of pretraining languages indeed leads to differences in zero-shot performance; (2) controlling for the amount of data allotted for each language during pretraining questions the primacy of English as the main *pretraining* language; and (3) our hypothesis regarding the sets of most- and least- contributing languages holds in both downstream tasks, and against two additional control groups.

2 Metrics for Pretraining-Aware Cross-Lingual Transfer

In this section, we extend existing metrics for zero-shot cross-lingual transfer to account for *pretraining* languages. Intuitively, our metrics for a model M and a given downstream task take into account three factors: (1) P, the set of languages seen during pretraining, (2) $s \in P$, the *source* language which is used for finetuning, and (3) $t \in P$, the *target* language, which is only seen during inference.

Formally, we adapt the formulation of Hu et al. (2020) to define a *pretraining-aware* bilingual zero-shot transfer score $\mathcal{Z}$ as:[3]

$$\mathcal{Z}(s \to t|P) := \varepsilon(M^{P,s}, t) \qquad (1)$$

Where $M^{P,l}$ is a model pretrained on the set of languages P and finetuned on downstream task instances in the language $l \in P$, and $\varepsilon(M, l)$ is an evaluation of model M on instances in language l in terms of the downstream metric, e.g., word label accuracy for part of speech tagging.

Following, we extend the definition to a set of downstream test languages $D \subseteq P$ to measure P's aggregated effect on zero-shot performance, by averaging over all bilingual transfer combinations in D:

$$\mathcal{Z}_P(D) = \sum_{l_1, l_2 \in D : l_1 \neq l_2} \frac{\mathcal{Z}(l_1 \to l_2|P)}{|D|^2} \qquad (2)$$

In following sections, we will use these metrics to evaluate how different choices for pretraining languages influence downstream performance.

[1] For example, English was X100 more likely to be sampled in mBERT's pretraining data than Icelandic.

[2] There are 2^n possible pretraining configurations taking into account inclusion and omission of every language.

[3] We opt not to normalize the score by the monolingual performance as done in Turc et al. (2021), as we do not want the monolingual performance to affect the score. Rather, we are interested in how the choice of pretraining languages directly contributes to the performance.

Language	Code	Family	Size [M chars]	
			Wiki	Sample
Piedmontese	pms	Indoeuropean	14	10
Irish	ga	Indoeuropean	38	10
Nepali	ne	Indoeuropean	78	10
Welsh	cy	Indoeuropean	85	10
Finnish	fi	Uralic	131	10
Armenian	hy	Indoeuropean	174	10
Burmese	my	Sino-Tibetian	229	10
Hindi	hi	Indoeuropean	473	10
Telugu	te	Dravidian	533	10
Tamil	ta	Dravidian	573	10
Korean	ko	Korean	756	10
Greek	el	Indoeuropean	906	10
Hungarian	hu	Uralic	962	10
Hebrew	he	Afroasiatic	1,261	10
Chinese	zh	Sino-Tibetian	1,546	10
Arabic	ar	Afroasiatic	1,695	10
Slovak	sv	Indoeuropean	1,744	10
Japanese	ja	Japonese	3,288	10
French	fr	Indoeuropean	4,958	10
German	de	Indoeuropean	6,141	10
Russian	ru	Indoeuropean	6,467	10
English	en	Indoeuropean	14,433	10

Table 1: The data size of the full wikipedia dump for the different languages in our study (in millions of characters) versus our fixed sized sampling of it. This exemplifies both the linguistic diversity as well as the variance in data sizes in the original Wikipedia corpus, often used for pretraining PLMs. In contrast, we create a balanced pretraining dataset by sampling 10M characters from all languages such that they conform to the smallest language portion in our set (Piedmontese).

3 Data Selection

In order to test the effect of pretraining languages on cross-lingual transfer, we seek to collect a pretraining dataset which satisfies two requirements.

First, we want the set of languages to be diverse to capture a wide range of linguistic features. Towards that end, we collect a set of 22 diverse languages from 9 language families, as listed in Table 1. These represent a wide variety of alphabets, linguistic and morphological features. We note that our approach can be readily extended to other languages beyond those present in this study.

Most crucially, we aim to balance the amount of data and control for its domain across languages, so that these factors would not confound our findings. Below we outline design choices we made towards this goal.

3.1 Data Balancing

To achieve a balanced dataset across our languages, we sample consecutive sentences from every language's corresponding Wikipedia dump (cleaned and tokenized using wikiextractor (Attardi, 2015)),[4] such that each language is represented by 10 million characters. This amount was chosen to align all languages to the lower-resource ones (e.g., Piedmontese or Irish) which comprise approximately 10mb of data. We choose to sample texts from Wikipedia as it consists of roughly similar encyclopedic domain across languages, and is widely used for training PLMs (Devlin et al., 2019).

Can we balance the amount of information across languages? We note that a possible confound in our study is that languages may encode different amounts of *information* in texts of similar character count. This may happen due to differences in the underlying texts, or in inherent language properties.[5] To estimate the amount of information in each our 10^7 character partitions, we tokenize each language partition l with a word-piece tokenizer, and look at the ratio between total number of tokens in l and the number of unique tokens in l, finding a good correlation across all our languages ($r = 0.73$), which may indicate that our dataset is indeed balanced in terms of information. Our intuition is that an imbalanced amount of information would lead the tokenizer to "invest" more tokens in some of the languages while neglecting the less informative ones. In our evaluation we revisit outliers in this metric and ensure that they do not correlate with better performance.

Is our sample representative of the full Wikipedia corpus in each language? Another concern may be that our sampled corpus per language is not indicative of the full corpus for that language, which may be much larger (see Table 1). To test this, we created three discrete length distributions. Two length distribution for sentences (in terms of words and tokens, where we use the same word-piece tokenizer as in our experiments from Section 4.1), and word length distribution in terms of characters. We then compared those three distributions between our sample and the full data using Earth Movers Distance. All means and standard deviations scored below 0.001, indicating that indeed all samples are similarly distributed to their respective full corpus in terms of these metrics.

[4] The latest version available on November 2021.

[5] For example logographic or abjad writing systems may be more condensed than other alphabets (Perfetti and Liu, 2005)

4 Pretrain Language Configuration

In this section we describe a method for estimating the effect that different pretrain language combinations in our dataset would have have on downstream zero-shot performance, by evaluating bilingual performance on the pretraining MLM task. This is required since exhaustively computing each of the possible language combinations is infeasible and grows exponentially with the number of examined languages (2^n combinations).

We begin by describing our experimental setup, hyperparameters and hardware configuration (Section 4.1). Then, in Section 4.2, we outline our estimation method, that also yields a complete graph structure over our languages, which is amenable for future exploration and analyses (Section 4.3). Then, in Section 4.4, we use the graph to formulate a set of downstream cross-lingual hypotheses regarding how different languages will affect zero-shot performance. In Section 5 we validate these hypotheses on two downstream tasks.

4.1 Experimental Setup

For all evaluations discussed below, we train a BERT model (Devlin et al., 2019) with 4 layers and 4 attention heads, an MLM task head, and an embedding space of size 512.[6] We train a single wordpiece tokenizer (Wu et al., 2016) on 44 languages consisting of the 22 languages elaborated above, and 22 additional languages listed in the Appendix, each with the same amount of data as described above. This was done so we can increase our analysis to more languages in the future. We train the models with a batch size of 8 samples, with sentences truncated to 128 tokens.

Each language model was trained up to 4 epochs. This was determined by examining the training loss on 6 diverse languages in our set and observing that they converge around 4 epochs. Masks were applied to the masked language modeling task according to the default setting, generating 15% mask tokens and 10% random tokens for each input sequence (Devlin et al., 2019).

All of our models were trained on a single GPU core using nvidia tesla M60, nvidia gtx 980, and RTX 2080Ti. The training time varied from 80 to 120 minutes, depending on the GPU and languages.

4.2 Building a Pretraining Language Graph

Intuitively, we measure masked language modelling (MLM) performance when pretraining on a pair of languages (l_1, l_2) as a proxy to the extent of how l_1 and l_2 contribute to one another in zero-shot cross-lingual transfer.

This methodology relies on two assumptions. First, we assume that the cross-lingual zero-shot performance as defined in Equation 2 is *monotonic*, i.e., that adding pretrain languages will increase the aggregated downstream performance. This is defined formally as:

$$\forall P' \subseteq P : \mathcal{Z}_P(D) \geq \mathcal{Z}_{P'}(D) \tag{3}$$

Second, we assume that MLM performance correlates with downstream task performance, which is often the assumption made when training PLMs to minimize perplexity (Peters et al., 2018; Devlin et al., 2019).

Bilingual MLM finetune score. Formally, for every language pair $s, t \in P$, we compute the following finetune score, $\mathcal{F}$:

$$\mathcal{F}(s \rightarrow t) := \frac{\varepsilon(M^{s,t}, t) - \varepsilon(M^t, t)}{\varepsilon(M^t, t)} \tag{4}$$

Where $M^{s,t}$ is a model pretrained on s, t, and ε is an intrinsic evaluation metric for MLM.[7] I.e., $\mathcal{F}(s, t)$ estimates how much the target language t "gains" in the MLM task from additional pretraining on the source language s compared to monolingual pretraining on t.

Figure 2 depicts a weighted adjacency matrix where coordinate (i, j) corresponds to $\mathcal{F}(l_i \rightarrow l_j)$. Furthermore, the same information can be conveyed in a complete directed weighted graph format, where each node represents a language, and edge (l_1, l_2) is weighted by $\mathcal{F}(l_1 \rightarrow l_2)$.

Language-Level Donation and Recipience. Next, for each language $l \in P$ we compute a *Donation* score, $\mathcal{D}$, as an aggregate over all of its finetune scores as a source language (i.e., how much it contributed to other languages), and similarly an *recipience* score, $\mathcal{R}$, by aggregating over all its finetune scores as a target language, to measure how much l is contributed to by other languages.

[6] We use the implementation provided by Hugging Face: https://huggingface.co/bert-base-uncased.

[7] We specifically use mean reciprocal rank (MRR), which correlates with perplexity.

Formally:

$$\mathcal{D}(l) := \sum_{t \in P, t \neq l} \mathcal{F}(l \rightarrow t) \quad (5)$$

$$\mathcal{R}(l) := \sum_{s \in P, s \neq l} \mathcal{F}(s \rightarrow l) \quad (6)$$

We depict both donation and recipience scores as aggregate row and column vectors in Figure 2.

Thus, based on the two assumptions above, our hypothesis is that the downstream cross-lingual transfer will be similar to the sum of recipience scores for all pretraining languages. Formally:

$$\mathcal{Z}_P(D) \approx \sum_{l \in P} \mathcal{R}(l) \quad (7)$$

Moreover, higher donation scores for languages in the pretrain set will result in higher scores in the downstream task. Formally:

$$\sum_{l \in P} \mathcal{D}(l) \geq \sum_{l \in P'} \mathcal{D}(l) \Rightarrow \mathcal{Z}_P \geq \mathcal{Z}_{P'} \quad (8)$$

4.3 Language Graph Analysis

Below we outline several interesting observations which can be drawn based on the pretrain language weighted adjacency matrix shown in Figure 2.

To allow further exploration of this data, we will make available a web-based extensible interactive exploration interface (see Figure 4 in the Appendix), which can surface how different linguistic features interact with these scores through various filters and visualizations.

Some language combinations are detrimental. Negative transfer scores are present in many of the target languages, e.g., between Korean (ko) and Arabic (ar), which means that initializing a language model for Arabic with weights learned for Korean hurts MLM performance on Arabic, compared with the Arabic monolingual baseline. This seems nontrivial as it may seem reasonable that initializing a model with an existing language model would result in at least the same performance as random initialization.

The bilingual relation is not symmetric. The finetune scores are not symmetric, i.e., there exists l_1, l_2 for which $\mathcal{F}(l_1 \rightarrow l_2) \neq \mathcal{F}(l_2 \rightarrow l_1)$. For example, for German and Finnish we get $0.51 = \mathcal{F}(fi \rightarrow de) \neq \mathcal{F}(de \rightarrow fi) = -0.24$, which means that Finnish initialization helps German MLM, while the opposite relation does not hold.

Finetuning as transfusion: mapping the linguistic blood-bank. Plotting the donation and recipience scores defined in Equations 5, 6 of each language l, as shown in Figure 3, gives rise to three-way language classification, loosely reminiscent of human blood types. For example, *O-type languages* (like Arabic or Finnish) generally contribute in all training configurations as *source* but do not get contributed to as *target*, while *AB+ type languages* (like German or Hindi) are positively influenced as *target* languages in all configurations but do not contribute to other languages as much when being the *source* language.

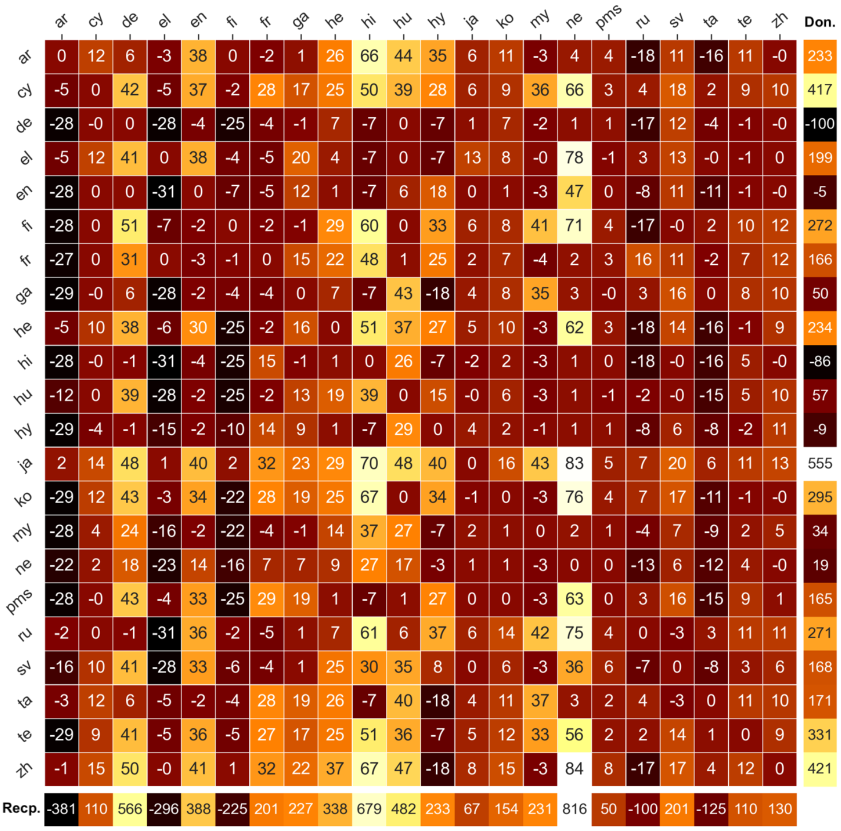
4.4 Choosing Pretrain Sets for Downstream

We use the donation scores to identify pretraining languages projected to lead to better downstream zero-shot performance, and the recipience score to identify downstream languages which will perform better as source languages for zero shot. This setup is summarized in Table 2.

Pretrain languages. We define three sets of languages that will be used for pretraining, using the donation score: (1) *Top donating:* Japanese, Telugu, Finnish, and Russian. This combination was chosen out of the top donating languages, while trying to keep the set linguistically diverse; (2) *Least donating:* Nepali, Burmese, Armenian, and English. This combination was chosen out of the least donating languages described above, while trying to keep the set linguistically diverse. We also chose to include English as it is a popular source language thanks to its abundance of data; and *Random:*A randomly selected set of 4 languages: Hebrew, Irish, French, and Slovak.

Downstream languages. To validate that lower recipience scores indeed means that languages are harder to improve via cross lingual transfer, we added 6 languages to all configurations described above — 3 *highly recipient* languages: Hindi, German, and Hungarian annotated R_h, and 3 relatively *least-recipient* languages: Arabic, Greek, and Tamil annotated R_l.

Hypotheses. We hypothesize that more donating pretraining sets will increase cross-lingual transfer in downstream tasks, and that mostly non-recipient languages will have lesser cross-lingual performance compared to recipient languages. These can be formally articulated using defintions from Equation 7 and 8:



	ar	cy	de	el	en	fi	fr	ga	he	hi	hu	hy	ja	ko	my	ne	pms	ru	sv	ta	te	zh	Don.
ar	0	12	6	-3	38	0	-2	1	26	66	44	35	6	11	-3	4	4	-18	11	-16	11	-0	233
cy	-5	0	42	-5	37	-2	28	17	25	50	39	28	6	9	36	66	3	4	18	2	9	10	417
de	-28	-0	0	-28	-4	-25	-4	-1	7	-7	0	-7	1	7	-2	1	1	-17	12	-4	-1	-0	-100
el	-5	12	41	0	38	-4	-5	20	4	-7	0	-7	13	8	-0	78	-1	3	13	-0	-1	0	199
en	-28	0	0	-31	0	-7	-5	12	1	-7	6	18	0	1	-3	47	0	-8	11	-11	-1	-0	-5
fi	-28	0	51	-7	-2	0	-2	-1	29	60	0	33	6	8	41	71	4	-17	-0	2	10	12	272
fr	-27	0	31	0	-3	-1	0	15	22	48	1	25	2	7	-4	2	3	16	11	-2	7	12	166
ga	-29	-0	6	-28	-2	-4	-4	0	7	-7	43	-18	4	8	35	3	-0	3	16	0	8	10	50
he	-5	10	38	-6	30	-25	-2	16	0	51	37	27	5	10	-3	62	3	-18	14	-16	-1	9	234
hi	-28	-0	-1	-31	-4	-25	15	-1	1	0	26	-7	-2	2	-3	1	0	-18	-0	-16	5	-0	-86
hu	-12	0	39	-28	-2	-25	-2	13	19	39	0	15	-0	6	-3	1	-1	-2	-0	-15	5	10	57
hy	-29	-4	-1	-15	-2	-10	14	9	1	-7	29	0	4	2	-1	1	1	-8	6	-8	-2	11	-9
ja	2	14	48	1	40	2	32	23	29	70	48	40	0	16	43	83	5	7	20	6	11	13	555
ko	-29	12	43	-3	34	-22	28	19	25	67	0	34	-1	0	-3	76	4	7	17	-11	-1	-0	295
my	-28	4	24	-16	-2	-22	-4	-1	14	37	27	-7	2	1	0	2	1	-4	7	-9	2	5	34
ne	-22	2	18	-23	14	-16	7	7	9	27	17	-3	1	1	-3	0	0	-13	6	-12	4	-0	19
pms	-28	-0	43	-4	33	-25	29	19	1	-7	1	27	0	0	-3	63	0	3	16	-15	9	1	165
ru	-2	0	-1	-31	36	-2	-5	1	7	61	6	37	6	14	42	75	4	0	-3	3	11	11	271
sv	-16	10	41	-28	33	-6	-4	1	25	30	35	8	0	6	-3	36	6	-7	0	-8	3	6	168
ta	-3	12	6	-5	-2	-4	28	19	26	-7	40	-18	4	11	37	3	2	4	-3	0	11	10	171
te	-29	9	41	-5	36	-5	27	17	25	51	36	-7	5	12	33	56	2	2	14	1	0	9	331
zh	-1	15	50	-0	41	1	32	22	37	67	47	-18	8	15	-3	84	8	-17	17	4	12	0	421
Recp.	-381	110	566	-296	388	-225	201	227	338	679	482	233	67	154	231	816	50	-100	201	-125	110	130	

Figure 2: Bilingual finetune scores between each pair of languages in our *balanced* corpus. Coordinate (i, j) represents $\mathcal{F}(l_i \rightarrow l_j)$, i.e., the performance in MRR[%] (which correlates with perplexity) of an LM pretrained on a bilingual corpus over languages (l_i, l_j) and tested intrinsically on l_j. The last column (marked *Don.*) sums over each line, i.e., index i in the column represents how much language i *donated* to all other languages. E.g., Japanese donated substantially, while German did not. Similarly, the j'th index in the last row (marked *Recp.*) sums over column j and represents how much language l_j improved in all configurations. E.g., Nepali improved significantly, while Arabic did not. We note that a dual representation of this heatmap is as a weighted adjacency matrix of a complete directed graph, where coordinate (i, j) represents the weight of the edge (l_i, l_j).

$$\forall P : \mathcal{Z}_P(R_h) > \mathcal{Z}_P(R_l) \tag{9}$$

$$\mathcal{Z}_{Top}(C) > \mathcal{Z}_{Random}(C) > \mathcal{Z}_{Bottom}(C) \tag{10}$$

5 Evaluating Downstream Zero-Shot Performance

All our pretrain configuration were trained in the same manner as in Section 4.1. We used the code and hyperparameter default values provided by XTREME to train the downstream tasks (Hu et al., 2020), adapted for multilinugal training and transfer.

5.1 Tasks

We evaluated all of the pretraining configurations detailed in Table 2 on two of XTREME's tasks: part of speech tagging (POS) and named entity recognition (NER). We aim to balance the data in both tasks across different finetune languages, so as not to skew results towards higher-resource languages.

For part-of-speech tagging, we use the data from the XTREME dataset, which in turn uses Universal Dependencies (Nivre et al., 2020). We truncated the data to 1000 sentences. For NER, we applied a similar procedure, where XTREME's data was taken from the Wikiann (panx) dataset (Rahimi et al., 2019).

5.2 Results

Several key observations can be made based on the results for both POS tagging and NER across all training configurations, which are presented in Tables 3 and 4. For each configuration P in *Top, Bottom, Random, Control* we calculated zeroshot transfer scores on the recipient set C using

	Base pretrain set		Shared pretrain set			Total data	Summary
	(Donating languages)		Most recipient (R_h)		Least recipient (R_l)		
Top	{ja, te, fi, ru}	+	{hi, de, hu}	+	{ar, el, ta}	10^8 characters	Most donating pretrain set.
Bottom	{ne, my, hy, en}		{hi, de, hu}		{ar, el, ta}	10^8 characters	Least donating pretrain set.
Random	{he, ga, fr, sv}		{hi, de, hu}		{ar, el, ta}	10^8 characters	Random donating pretrain set.
Control	{}		{hi, de, hu}		{ar, el, ta}	10^8 characters	No additional donating languages.

Table 2: Pretraining sets for for four language configuration. Each pretrain set is comprised of *donating* languages and *recipient* languages. Notice that the control group has less languages but the same amount of data overall, equally distributed among its pretraining languages.

	NER $[F_1]$		**POS** $[F_1]$	
	Avg. Monolingual	Avg. Zeroshot	Avg. Monolingual	Avg. Zeroshot
Top	**0.570**	**0.175**	**0.615**	**0.27**
Random	0.291	0.123	0.611	0.26
Bottom	0.323	0.119	0.605	**0.27**
Control	0.331	0.119	**0.623**	**0.27**

Table 3: Donation results for named entity recognition and part of speech tagging. For each pretraining language group (**Top**, **Random**, **Bottom**, and **Control**), we report their corresponding average monolingual and zero shot performance. **Top** consistently outperforms other configuration on NER, while results are more similar in POS, where all of the configurations are roughly on par for zero-shot performance.

	NER $[F_1]$		**POS** $[F_1]$	
	Avg. Monolingual	Avg. Zeroshot	Avg. Monolingual	Avg. Zeroshot
Most recipient (R_h)	**0.390**	**0.162**	**0.640**	**0.285**
Least recipient (R_l)	0.387	0.106	0.588	0.253

Table 4: Recipience results. We report aggregate results across different training configurations for two groups of downstream recipient languages. As we predict from intrinsic pretraining performance, the **Most Recipient** set does better than the **Least recipient** set across both tasks in zero-shot and monolingual performance.

$\mathcal{Z}_P(C)$ defined by Equation 2. Monolingual results under each pretrain set P were calculated by the average $F1$ performance of each language in C: $\frac{1}{|C|}\sum_{l \in C} \varepsilon(M^{P,l}, l)$ where $\varepsilon(M^{P,l}, l)$ denotes the $F1$ score of model pretrained on P, finetuned on l and evaluated on l.

Pretraining configuration affects downstream cross-lingual transfer results. In both tasks, we observe a variance of results when changing the pretraining configuration, despite all of the configuration having similar amounts of data. This effect is more pronounced in NER, where the average monolingual performance varies by roughly 28% between the best and least performing configuration, and zero-shot performance varies by more than 5%. This shows that previous work in analyzing cross-lingual performance has indeed omitted an important interfering factor.

Recipient score correlates with downstream cross-lingual performance. We evaluated zero-shot transfer scores for each recipient language set $R \in \{R_l, R_h\}$ as the average zero-shot transfer scores over all pretrain configurations. Table 4 reveals that in all tasks and configurations the more recipient set scored better zero-shot scores compared to the least recipient set, i.e. exhibit better cross lingual performance (+5% in NER, +3% in POS tagging).

Multilingual pretraining can improve *monolingual* performance. As seen in Table 3, the *Top* pretrain configuration scored a monolimgual score higher than the control group by 23%. This suggests that multilingual pretraining datasets can benefit monolingual downstream results even more than datasets with the same amount of data spread across less languages.

English might not be an optimal pretrain language. Corresponding with our previous results, if *donation score* is indicative to a languages contribution in pretraining, English's relative low donation score might indicate that it is not the best language to pretrain upon. English was also part of the *bottom* pretrain configuration which scored lower than *top* as seen in Table 3. Further research

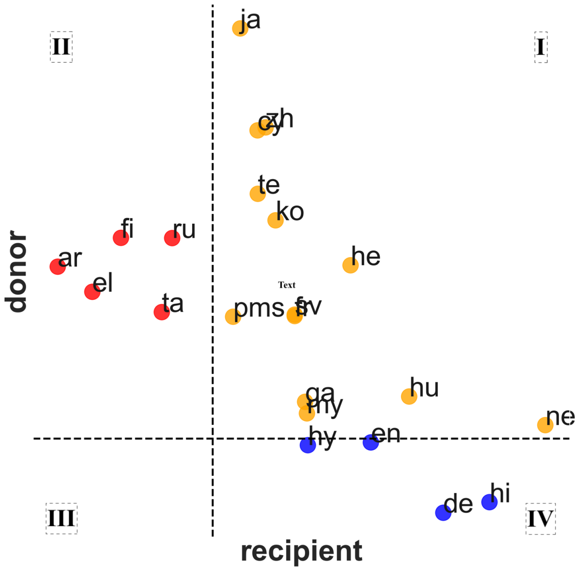


Figure 3: Our languages on a "donor" versus "recipient" graph. A positive coordinate on the "donor" score (X axis) represents a language which on average improved other language's performance in bilingual pretraining, while a negative score indicates a language which hurt other languages on average. Inversely, positive score on the Y axis represents languages whose performance was improved by bilingual pretraining, while negative scores represent language whose performance was hurt by it. The II quadrant represents O type languages (donating but not receiving), languages on the IV's quadrant are AB+ type languages (receiving but not donating).

and broader analysis are required to ascertain this finding.

6 Limitations and Future Work

As other works on cross-lingual transfer, our results are influenced by many hyperparameters. Below we explicitly define our design choices and how they can be explored in future work.

First, data scarcity in low-resource languages restricted us to small data amounts. Although our experiments showed non-trivial signal for pretraining and downstream tasks, future work may apply our framework to larger data sizes.

Second, for efficiency sake, we trained relatively small models to enable us to train a large amount of language configurations, while ensuring convergence in 6 languages. Furthermore we did not do any hyper-parameter tuning and use only values reported in previous work. Moreover, we limit ourselves to BERTs architecture. Future work may revisit any of these design choices to shed more light on their effect.

Finally, our downstream analysis focused on POS tagging and NER since they were available for many languages. Further experimentation can test if our results hold for more NLP tasks.

7 Related Work

To the best of our knowledge, we are the first work to focus and control for the amount of data allocated for each language during pretraining and fine-tuning. Other works have explored cross-lingual transfer from different angles.

Perhaps most related to our work, Turc et al. (2021) challenge the primacy of English as a source language for cross-lingual transfer in various downstream tasks. Their work shows that German and Russian are often more effective sources. In all of their experiments they use mBERT's imbalanced pretraining corpus.

Wu and Dredze (2020) evaluate how mBERT performs on a wide set of languages, focusing on the quality of representation for low-resource languages in various downstream tasks by defining a scale from low to high resource. They show that mBERT underperforms non BERT monolingual baselines for low resource languages while performing well for high resource ones.

Finally, Conneau et al. (2019) introduce the transfer-interference trade-off where low resource languages benefit from scaling to more languages, until a given point where the overall performance on monolingual and cross-lingual benchmarks degrades.

8 Conclusions

We explored the effect of pretraining language selection on downstream zero-shot transfer. We first curate a balanced set of pretraining languages. Second, we devise a novel estimation technique, quadratic in the number of languages, which projects which pretraining languages will serve better in cross-lingual downstream transfer and which specific downstream languages will do best in that setting. Finally, we test our hypothesis on two downstream multilignual tasks, and show that pretraining language selection choice indeed leads to varying downstream cross lingual results, and that our estimation is a good technique for projecting performance. Taken together our results suggest that pretraining language selection should be a factor in estimating cross-lingual transfer, and that current practices which focus on high-resource languages may be sub-optimal.

9 Broader Impact and Ethical Considerations

Our work did not involve any new data or annotation collection, and as such did not require crowd-sourced or in-house workers, or introduced any new models and related risks. Instead, we examine different combinations of multilingual pretraining, which may lead to better multilingual models in a vast array of languages, especially in low-resource or zero-shot settings.

References

Giuseppe Attardi. 2015. Wikiextractor. `https://github.com/attardi/wikiextractor`.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Koren Lazar, Benny Saret, Asaf Yehudai, Wayne Horowitz, Nathan Wasserman, and Gabriel Stanovsky. 2021. Filling the gaps in Ancient Akkadian texts: A masked language modelling approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4682–4691, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.

Charles A Perfetti and Ying Liu. 2005. Orthography to phonology and meaning: Comparisons across and within writing systems. *Reading and Writing*, 18(3):193–210.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for ner. *arXiv preprint arXiv:1902.00193*.

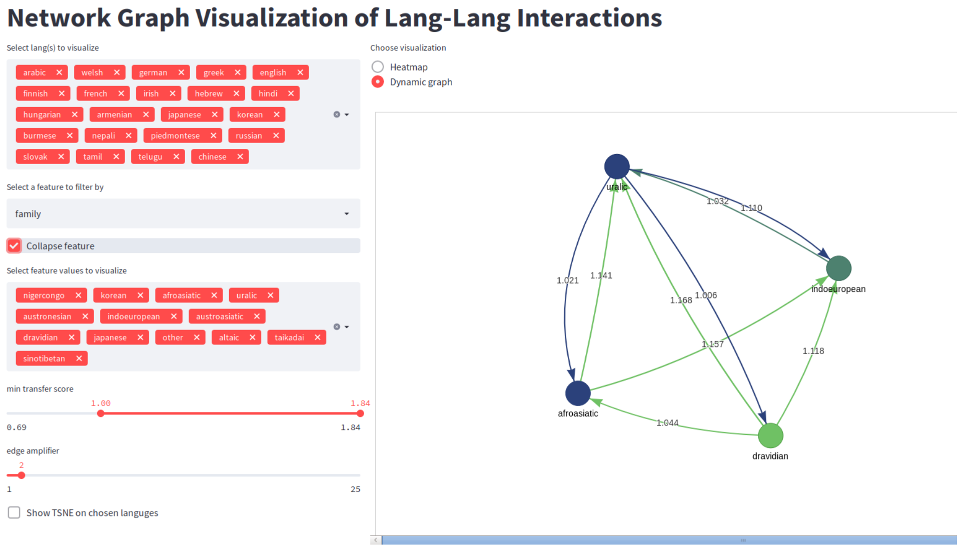
Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *arXiv preprint arXiv:2106.16171*.

Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Emerging cross-lingual structure in pretrained language models. *arXiv preprint arXiv:1911.01464*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.



Network Graph Visualization of Lang-Lang Interactions

Select lang(s) to visualize

arabic ✕ welsh ✕ german ✕ greek ✕ english ✕ finnish ✕ french ✕ irish ✕ hebrew ✕ hindi ✕ hungarian ✕ armenian ✕ japanese ✕ korean ✕ burmese ✕ nepali ✕ piedmontese ✕ russian ✕ slovak ✕ tamil ✕ telugu ✕ chinese ✕

Select a feature to filter by

family

☑ Collapse feature

Select feature values to visualize

nigercongo ✕ korean ✕ afroasiatic ✕ uralic ✕ austronesian ✕ indoeuropean ✕ austroasiatic ✕ dravidian ✕ japanese ✕ other ✕ altaic ✕ taikadai ✕ sinotibetan ✕

min transfer score

1.00 1.84

0.69 1.84

edge amplifier

2

1 25

□ Show TSNE on chosen langugues

Choose visualization

○ Heatmap

● Dynamic graph

Figure 4: Our visualization tool, based on Streamlit (https://streamlit.io)

A Appendix

Full list of tokenized languages The full list of wikipedia language codes for languages used in our tokenizer training is: af, am, ar, ca, cs, cy, da, de, el, en, es, fa, fi, fr, ga, he, hi, hu, hy, id, is, it, ja, ko, mg, my, ne, nl, pl, pms, ro, ru, sk, sv, sw, ta, te, th, tr, ur, vi, yi, zh.

Visualization tool Figure 4 presents our interactive visualization tool for the finetune experiments reported in Section 4. We allow visualizing MLM fine-tune score as a heatmap or as a dynamic directed graph. The graph is based on the adjacency matrix induced by our models as described in Section 4. The adjacency matrix can be manipulated using a controlled threshold to dilute edges. Each feature can be visualized on the graph in two ways. One way is to color all nodes (languages) w.r.t to their feature value, another way is to collapse all feature instances (languages) into a single averaged feature node whose out-coming edges are the averages over all nodes tagged by this feature, as seen in the figure.