When is BERT Multilingual? Isolating Crucial Ingredients for Cross-lingual Transfer

Anonymous ACL submission

Abstract

While recent work on multilingual language models has demonstrated their capacity for cross-lingual zero-shot transfer on downstream tasks, there is a lack of consensus in the community as to what shared properties between languages enable such transfer. Analyses involving pairs of natural languages are often inconclusive and contradictory since languages simultaneously differ in many linguistic aspects. In this paper, we perform a large-scale empirical study to isolate the effects of various linguistic properties by measuring zero-shot transfer between four diverse natural languages and their counterparts constructed by modifying aspects such as the script, word order, and syntax. Among other things, our experiments show that the absence of sub-word overlap significantly affects zero-shot transfer when languages differ in their word order, and there is a strong correlation between transfer performance and word embedding alignment between languages (e.g., $\rho_s = 0.94$ on the task of NLI). Our results call for focus in multilingual models on explicitly improving word embedding alignment between languages rather than relying on its implicit emergence. [1]

1 Introduction

Multilingual language models like XLM (Conneau et al., 2020a) and Multilingual-BERT[2] are trained with masked-language modeling (MLM) objective on a combination of raw text from multiple languages. Surprisingly, these models exhibit decent cross-lingual zero-shot transfer, where fine-tuning on a task in a source language translates to good performance for a different language (target).

Requirements for zero-shot transfer Recent studies have provided inconsistent explanations for properties required for zero-shot transfer (hereon, transfer). For example, while Wu and Dredze (2019) conclude that sub-word overlap is vital for transfer, K et al. (2020) demonstrate that it is not crucial, although they consider only English as the source language. While Pires et al. (2019) suggest that typological similarity (e.g., similar SVO order) is essential for transfer, other works (Kakwani et al., 2020; Conneau et al., 2020a) successfully build multilingual models for dissimilar languages.

Need for systematic analysis A major cause of these discrepancies is a large number of varying properties (e.g., syntax, script, and vocabulary size) between languages, which make isolating crucial ingredients for transfer difficult. Some studies alleviate this issue by creating synthetic languages which differ from natural ones only in specific linguistic properties like script (K et al., 2020; Dufter and Schütze, 2020). However, their focus is only on English as a source language, and the scale of their experiments is small (in number of tasks or pre-training corpora size), thus limiting the scope of their findings to their settings alone.

Our approach We perform a systematic study of cross-lingual transfer on bilingual language models trained on a natural language and a systematically *derived* counterpart. We choose four diverse natural languages (English, French, Arabic, and Hindi) and create *derived* variants using four different transformations on structural properties such as inverting or permuting word order, altering scripts, or varying syntax (Section 3.2). We train models on each of the resulting sixteen language pairs, and evaluate zero-shot transfer on four downstream tasks – natural language inference (NLI), named-entity recognition (NER), part-of-speech tagging (POS), and question-answering (QA).

Our experiments show that:

1. Contrary to previous belief, the absence of sub-word overlap degrades transfer when languages

[1] Code: Provided in the supplementary material.

[2] https://github.com/google-research/bert/blob/master/multilingual.md

differ in their word order (e.g., by more than 40 F1 points on POS tagging, (§ 4.1)).

2. There is a strong correlation between token embedding alignment and zero-shot transfer across different tasks (e.g., $\rho_s = 0.94, p < .005$ for XNLI, Fig 4).
3. Using pre-training corpora from similar sources for different languages (e.g., Wikipedia) boosts transfer when compared to corpora from different sources (e.g., 17 F1 points on NER, Fig 3).

To our knowledge, we are the first study to quantitatively show that zero-shot transfer between languages is strongly correlated with token embedding alignment ($\rho_s = 0.94$ for NLI). We also show that the current multilingual pre-training methods (Conneau et al., 2020a; Doddapaneni et al., 2021) fall short of aligning embeddings even between simple natural and derived language pairs, leading to failure in zero-shot transfer. Our results call for training objectives that explicitly improve alignment using either supervised (e.g., parallel corpora and bilingual dictionaries) or unsupervised data.

2 Related work

Multilingual pre-training for Transformers The success of monolingual Transformer language models (Devlin et al., 2019; Radford et al., 2018) has driven studies that learn a multilingual language-model (LM) on several languages. Multilingual-BERT[3] (M-BERT) is a single neural network pre-trained using the masked language-modeling (MLM) objective on a corpus of text from 104 languages. XLM (Conneau and Lample, 2019) introduced translation language-modeling, which performs MLM on pairs of parallel sentences, thus encouraging alignment between their representations. These models exhibit surprising zero-shot cross-lingual transfer performance (Conneau and Lample, 2019; K et al., 2020), a setup where the model is fine-tuned on a source language and evaluated on a different target language.

Analysis of cross-lingual transfer While Pires et al. (2019), Conneau et al. (2020b), and K et al. (2020) showed that transfer works even without a shared vocabulary between languages, Wu and Dredze (2019) discovered a correlation between sub-word overlap and zero-shot performance. Conneau et al. (2020b) and Artetxe et al. (2020a) showed that shared parameters for languages with different scripts were crucial for transfer.

Pires et al. (2019) and (Wu and Dredze, 2019) observed that transfer for NER and POS tagging works better between typologically similar languages. However, a study conducted by Lin et al. (2019) showed that there is no simple rule of thumb to gauge when transfer works between languages.

Transfer between real and synthetic Languages K et al. (2020) create a synthetic language by changing English's script and find that transfer between it and Spanish works even without common sub-words. However, they use only English as their source language, test only on two tasks, and use a single natural-synthetic language pair. Dufter and Schütze (2020) study transfer between English and *synthetic* English obtained by changing the script, word order, or model delimiters. However, they use a small corpus (228K words) compared to current standards (we use 3 orders more) and measure only embedding similarity and not zero-shot transfer.

3 Approach

We first provide some background on bilingual language models (Section 3.1), followed by descriptions of our transformations (Section 3.2), and our training and evaluation setup (Section 3.3).

3.1 Background

Bilingual pre-training The standard setup (Conneau and Lample, 2019) trains a bilingual language model (*Bi-LM*) on raw text corpora from two languages simultaneously. *Bi-LM* uses the masked language-modeling loss ($\mathcal{L}_{\text{MLM}}$) on the corpora from the two languages ($\mathcal{C}_1, \mathcal{C}_2$) separately with no explicit cross-lingual signal:

$$\mathcal{L}^{\theta}_{\text{Bi-LM}}(\mathcal{C}_1 + \mathcal{C}_2) = \mathcal{L}^{\theta}_{\text{MLM}}(\mathcal{C}_1) + \mathcal{L}^{\theta}_{\text{MLM}}(\mathcal{C}_2)$$

A shared byte pair encoding tokenizer (Sennrich et al., 2015) is trained on $\mathcal{C}_1 + \mathcal{C}_2$. A single batch contains instances from both languages, but each instance belongs to a single language.

Zero-shot transfer evaluation Consider a bilingual model (*Bi-LM*) pre-trained on two languages, *source* and *target*. Zero-shot transfer involves fine-tuning *Bi-LM* on downstream task data from *source* and evaluating on test data from *target*. This is considered zero-shot because *Bi-LM* is not fine-tuned on any data belonging to *target*.

[3]https://github.com/google-research/bert/blob/master/multilingual.md

Transformation	**Instance (s)**	**Transformed instance ($\mathcal{T}(s)$)**
Inversion ($\mathcal{T}_{\mathbf{inv}}$)	Welcome to NAACL at Seattle	Seattle at NAACL to Welcome
Permutation ($\mathcal{T}_{\mathbf{perm}}$)	This is a conference	a This conference is
Transliteration ($\mathcal{T}_{\mathbf{trans}}$)	I am Sam . I am	♣(I) ♡(am) ♢(Sam) ♠(.) ♣(I) ♡(am)
Syntax ($\mathcal{T}_{\mathbf{syn}}$)	Sara (S) ate (V) apples (O)	Sara (S) apples (O) ate (V)
	Une table (N) ronde (A)	Une ronde (A) table (N)

Table 1: Examples of our transformations applied to different sentences (without sub-word tokenization). *Inversion* inverts the tokens, *Permutation* samples a random reordering, and *Transliteration* changes the script. We use symbols (♣) to denote words in the new script and mention the corresponding original word in brackets. *Syntax* stochastically modifies the syntactic structure. In the first example for *Syntax*, the sentence in Subject-Verb-Object (SVO) order gets transformed to SOV order, and in the second, the sentence in Noun-Adjective (NA) order gets transformed to the AN order. The examples are high probability re-orderings and other ones might be sampled too.

3.2 Generating language variants with systematic transformations

Natural languages typically differ in several ways, like the script, word order, and syntax. To isolate the affect of these properties on zero-shot transfer, we obtain *derived* language corpora (hereon, *derived* corpora) from *original* (natural) language corpora by performing sentence level transformations ($\mathcal{T}$) which change particular properties. For example, an "*inversion*" transformation could be used to invert each sentence in the corpus ($Welcome_1$ to_2 $NAACL_3 \Rightarrow NAACL_3$ to_2 $Welcome_1$). Since the transformation ($\mathcal{T}$) is applied on each sentence of the *original* corpus, the size of the *original* and the *derived* corpus is the same. In the following sections, we will use the following notation:

$$\begin{aligned}\mathcal{C}_{\mathbf{orig}} &\equiv \text{Original corpus}\\ &= \{s_i \mid i = 1:N, s_i = \text{sentence}\}\\ \mathcal{T} &\equiv \text{Sentence-level transformation}\\ \mathcal{C}_{\mathbf{deriv}} &\equiv \text{Derived corpus}\\ &= \{\mathcal{T}(\text{sent}) \mid \forall\ \text{sent} \in \mathcal{C}_{\mathbf{orig}}\}\end{aligned}$$

Types of transformations We consider four transformations which modify different aspects of sentences (examples in Table 1):

1. **Inversion** ($\mathcal{T}_{\mathbf{inv}}$): Invert the order of *tokens* in the sentence, like in Dufter and Schütze (2020). The first token becomes the last, and vice versa.
2. **Permutation** ($\mathcal{T}_{\mathbf{perm}}$): Permute the order of tokens in a sentence uniformly at random. For a sentence of n tokens, we sample a random ordering with probability $\frac{1}{n!}$.
3. **Transliteration** ($\mathcal{T}_{\mathbf{trans}}$): Change the script of all tokens other than the special tokens (like `[CLS]`). This creates a *derived* vocabulary ($\mathcal{V}_{deriv}$) with a one-to-one correspondence with the original vocabulary ($\mathcal{V}_{orig}$).
4. **Syntax** ($\mathcal{T}_{\mathbf{syn}}$): Modify a sentence to match the syntactic properties of a different natural language by re-ordering the dependents of nouns and verbs in the dependency parse. These transformations are stochastic because of the errors in parsing and sampling over possible re-orderings (Wang and Eisner, 2016).

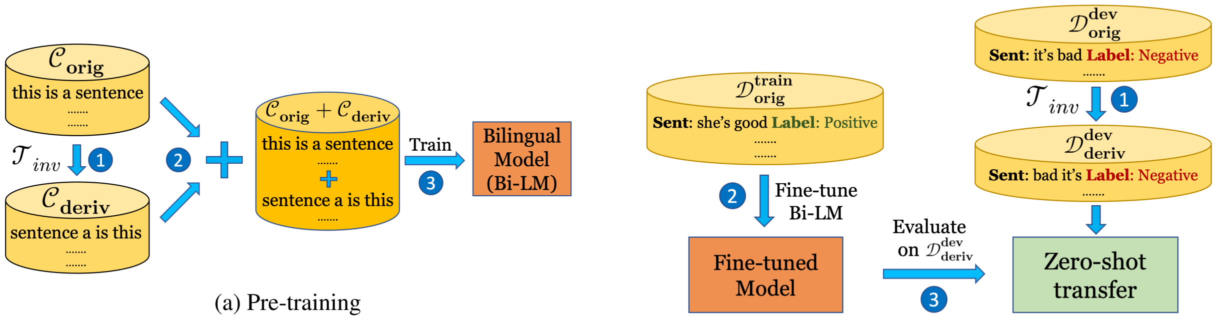
Transformations for downstream tasks We obtain the downstream corpus in the *derived* language ($\mathcal{D}_{\mathbf{deriv}}$) by applying the same transformation ($\mathcal{T}$) used during pre-training on the *original* downstream corpus ($\mathcal{D}_{\mathbf{orig}}$). Unlike pre-training corpora which contain raw sentences, instances in downstream tasks contain one or more sentences with annotated labels. For text classification tasks like NLI, we apply the transformation on each sentence in every dataset instance. For token classification tasks (e.g., NER, POS), any transformation which changes the order of the tokens also changes the order of the labels. We present the mathematical specification in Appendix A.

3.3 Model Training and Evaluation

We now describe our pre-training and zero-shot transfer evaluation setup. Figure 1 provides an overview of pre-training and fine-tuning, and Table 2 summarizes the evaluation metrics we use.

Pre-training Let $\mathcal{C}_{\mathbf{orig}}$ and $\mathcal{C}_{\mathbf{deriv}}$ be the *original* and *derived* language pre-training corpora. We train two models for each *original-derived* pair:

1. **Bilingual Model (*Bi-LM*)**: A bilingual model pre-trained on the combined corpus ($\mathcal{C}_{\mathbf{orig}}+\mathcal{C}_{\mathbf{deriv}}$) (Figure 1a).



(a) Pre-training

(b) Fine-tuning

Figure 1: **(a)** During pre-training, we 1 obtain the *derived* language corpus ($\mathcal{C}_{\text{deriv}}$) by *transforming* the *original* language corpus ($\mathcal{C}_{\text{orig}}$). 2 The two corpora are combined and, 3 a bilingual model (*Bi-LM*) is learned using the MLM objective. **(b)** During fine-tuning, we 1 obtain the derived dev dataset ($\mathcal{D}^{\text{dev}}_{\text{deriv}}$) by transforming the original dev dataset ($\mathcal{D}^{\text{dev}}_{\text{orig}}$). 2 *Bi-LM* is fine-tuned on the original train dataset ($\mathcal{D}^{\text{train}}_{\text{orig}}$), and 3 evaluated on $\mathcal{D}^{\text{dev}}_{\text{deriv}}$, which is the standard zero-shot cross lingual setup.

Evaluation	Corpus source		
	Pre-train	Fine-tune (train)	Fine-tune (dev)
BZ	$\mathcal{C}_{\text{orig}} + \mathcal{C}_{\text{deriv}}$	$\mathcal{D}_{\text{orig}}$	$\mathcal{D}_{\text{deriv}}$
BS	$\mathcal{C}_{\text{orig}} + \mathcal{C}_{\text{deriv}}$	$\mathcal{D}_{\text{deriv}}$	$\mathcal{D}_{\text{deriv}}$
MZ	$\mathcal{C}_{\text{orig}}$	$\mathcal{D}_{\text{orig}}$	$\mathcal{D}_{\text{deriv}}$
$\Delta_{(\text{BZ}-\text{BS})} = (\textbf{BZ} - \textbf{BS})$			
$\Delta_{(\text{MZ}-\text{BS})} = (\textbf{MZ} - \textbf{BS})$			

Table 2: Summary of evaluation metrics defined in § 3.3. $\mathcal{C}$ and $\mathcal{D}$ denote the pre-training and downstream corpus respectively, and their subscript indicates their source (*original* or *derived*). **BZ** and **MZ** represent bilingual and monolingual zero-shot transfer scores, and **BS** is the supervised learning baseline on *derived*. The differences in the setting of **BZ** and other scores are typeset in blue. We use $\Delta_{(\mathbf{BZ-BS})}$ and $\Delta_{(\mathbf{MZ-BS})}$ (defined in the last two rows) throughout our paper.

2. **Monolingual Model (*Mono-LM*)**: A monolingual model trained only on $\mathcal{C}_{\text{orig}}$ for the same number of steps as *Bi-LM*'s. *Mono-LM* is used as a baseline to measure zero-shot transfer of a model not pre-trained on *derived*.

Evaluation Let $\mathcal{D}^{\text{train}}_{\text{orig}}$ and $\mathcal{D}^{\text{dev}}_{\text{orig}}$ be the *original* language training and development sets for a downstream task, and $\mathcal{D}^{\text{train}}_{\text{deriv}}$ and $\mathcal{D}^{\text{dev}}_{\text{deriv}}$ be the corresponding *derived* language datasets. For evaluation, we first fine-tune the pre-trained models on a downstream training set and evaluate the resulting model on a development set (Figure 1b). Since our goal is to investigate the extent of zero-shot transfer, we require appropriate lower and upper bounds to make informed conclusions. To this end, we compute three metrics, all on the same development set (summarized in Table 2):

- **Bilingual zero-shot transfer (*BZ*)**: This is the standard zero-shot transfer score (Conneau and Lample, 2019) which measures how well a bilingual model fine-tuned on $\mathcal{D}^{\text{train}}_{\text{orig}}$ zero-shot transfers to the other language ($\mathcal{D}^{\text{dev}}_{\text{deriv}}$).
- **Bilingual supervised synthetic (*BS*)**: This is the supervised learning performance on the *derived* language obtained by fine-tuning *Bi-LM* on $\mathcal{D}^{\text{train}}_{\text{deriv}}$ and evaluating it on $\mathcal{D}^{\text{dev}}_{\text{deriv}}$.
- **Monolingual zero-shot transfer (*MZ*)**: This measures the zero-shot performance of the baseline *Mono-LM*, which is not pre-trained on the *derived* language, by fine-tuning *Mono-LM* on $\mathcal{D}^{\text{train}}_{\text{orig}}$ and evaluating it on $\mathcal{D}^{\text{dev}}_{\text{deriv}}$.

BS uses fine-tuning train data from the *derived* language and serves as an upper-bound on *BZ* and *MZ* which don't use it. *MZ* doesn't pre-train on the *derived* language and serves as a lower-bound on *BZ* which does pre-train on it. For easier comparison of *BZ* and *MZ* with *BS* (upper-bound), we report the following score differences (Table 2), which are both negative in our experiments.

$$\Delta_{(\text{BZ}-\text{BS})} = (\text{BZ} - \text{BS}) \tag{1}$$

$$\Delta_{(\text{MZ}-\text{BS})} = (\text{MZ} - \text{BS}) \tag{2}$$

BZ alone cannot capture the quality of the zero-shot transfer. A large and negative $\Delta_{(\text{BZ}-\text{BS})}$ implies that bilingual zero-shot transfer is much worse than supervised fine-tuning on *derived*. Concurrently, $\Delta_{(\text{BZ}-\text{BS})} \approx \Delta_{(\text{MZ}-\text{BS})}$ implies that *Bi-LM* transfers as poorly as *Mono-LM*. **Thus, good zero-shot transfer is characterized by $\Delta_{(\mathbf{BZ-BS})} \approx 0$ and $\Delta_{(\mathbf{BZ-BS})} \gg \Delta_{(\mathbf{MZ-BS})}$.**

Task	Inversion ($\mathcal{T}_{\mathbf{inv}}$)			Permutation ($\mathcal{T}_{\mathbf{perm}}$)			Syntax ($\mathcal{T}_{\mathbf{syn}}$)			Transliteration ($\mathcal{T}_{\mathbf{trans}}$)		
	$\Delta_{(\mathrm{BZ-BS})}$	$\Delta_{(\mathrm{MZ-BS})}$	*BZ*	$\Delta_{(\mathrm{BZ-BS})}$	$\Delta_{(\mathrm{MZ-BS})}$	*BZ*	$\Delta_{(\mathrm{BZ-BS})}$	$\Delta_{(\mathrm{MZ-BS})}$	*BZ*	$\Delta_{(\mathrm{BZ-BS})}$	$\Delta_{(\mathrm{MZ-BS})}$	*BZ*
XNLI	-10.2	-13.0	58.4	-3.6	-8.6	62.6	-0.9 *	-1.1	67.8	-1.0 *	-36.7	69.3
NER	-49.1	-46.7	37.9	-26.3	-35.4	47.3	-14.6	-16.6	62.9	-1.9 *	-82.6	83.7
POS	-30.2	-36.2	64.2	-11.2	-25.2	73.6	-4.4	-7.6	89.4	-0.4 *	-95.0	95.4
XQuAD[4]	-32.8	-31.0	22.8	—[4]	—	—	—[4]	—	—	0.0 *	-55.9	61.2

Table 3: **(1) Evaluation:** We report $\Delta_{(\mathbf{BZ-BS})}$ and $\Delta_{(\mathbf{MZ-BS})}$ (§ 3.3 and Table 2) for transformations on different tasks, averaged over four languages (EN, FR, HI, AR). We report the breakdown for different languages in Appendix B. *BZ*, the bilingual zero-shot performance, is reported for reference. **(2) Interpreting scores:** Smaller (more negative) $\Delta_{(\mathrm{BZ-BS})}$ implies worse bilingual zero-shot transfer, whereas $\Delta_{(\mathrm{BZ-BS})} \approx 0$ implies strong transfer. $\Delta_{(\mathrm{BZ-BS})} \gg \Delta_{(\mathrm{MZ-BS})}$ implies that bilingual pre-training is extremely useful. Scores are highlighted based on their value (lower scores have a higher intensity of red). Cases with strong zero-shot transfer ($\Delta_{(\mathrm{BZ-BS})} \approx 0$) are marked with an asterisk. **(3) Trends:** $\mathcal{T}_{\mathbf{trans}}$ exhibits strong transfer on all tasks and languages (high $\Delta_{(\mathrm{BZ-BS})}$ scores), and bilingual pre-training is extremely useful ($\Delta_{(\mathrm{BZ-BS})} \gg \Delta_{(\mathrm{MZ-BS})}$), implying that zero-shot transfer is possible between languages with different scripts but the same word order. $\mathcal{T}_{\mathbf{inv}}$ and $\mathcal{T}_{\mathbf{perm}}$ suffer on all tasks (small $\Delta_{(\mathrm{BZ-BS})}$ scores) whereas $\mathcal{T}_{\mathbf{syn}}$ suffers significantly lesser, which provides evidence that local changes to the word order made by *Syntax* ($\mathcal{T}_{\mathbf{syn}}$) hurts zero-shot transfer significantly lesser than global changes made by *Inversion* ($\mathcal{T}_{\mathbf{inv}}$) and *Permutation* ($\mathcal{T}_{\mathbf{perm}}$).

Dataset	Task	Metric
XNLI (Conneau et al., 2018)	NLI	Accuracy
Wikiann (Pan et al., 2017)	NER	F1
UD v2.5 (Nivre et al., 2018)	POS	F1
XQuAD (Artetxe et al., 2020b)	QA	F1

Table 4: XTREME benchmark datasets used for zero-shot transfer evaluation. NLI=Natural Language Inference, NER=Named-entity recognition, POS=Part-of-speech tagging, QA=Question-Answering.

3.4 Experimental Setup

Languages We choose four diverse natural languages: English (Indo-European, Germanic), French (Indo-European, Romance), Hindi (Indo-European, Indo-Iranian), and Arabic (Afro-Asiatic, Semitic), which are represented in the multilingual XTREME benchmark (Hu et al., 2020). For each language, we consider four transformations (Section 3.2) to create *derived* counterparts, giving us 16 different original-derived pairs in total. For the *Syntax* transformation, we use Qi et al. (2020) for parsing. We modify the syntax of FR, HI, and AR to that of EN, and the syntax of EN to that of FR.

Datasets For the pre-training corpus ($\mathcal{C}_{\mathbf{orig}}$), we use a 500MB (uncompressed) subset of Wikipedia ($\approx$ 100M tokens) for each language. This matches the size of WikiText-103 (Merity et al., 2016), a standard language-modeling dataset. For downstream evaluation, we choose four tasks from the XTREME benchmark (Hu et al., 2020). Table 4 lists all the datasets and their evaluation metrics.

Implementation Details We use a variant of RoBERTa (Liu et al., 2019) which has 8 layers, 8 heads, and a hidden dimensionality of 512. We train each model on 500K steps, a batch size of 128, and a learning rate of $1e\text{-}4$ with a linear warmup of 10K steps. We use an *original* language vocabulary size of 40000 for all the models and train on 8 Cloud TPU v3 cores for 32-48 hours. For fine-tuning, we use standard hyperparameters (Appendix F) from the XTREME benchmark and report our scores on the development sets.

4 Results

Our experiments reveal several interesting findings for bilingual models including the situational importance of sub-word overlap for zero-shot transfer (§ 4.1, 4.2), the effect of domain mismatch between languages (§ 4.3), and correlation of zero-shot performance with embedding alignment (§ 4.4). We connect our findings to zero-shot transfer results between natural languages in Section 4.5.

[4] XQuAD is a question-answering task where the correct answer is a *contiguous* span. We do not report scores on XQuAD for $\mathcal{T}_{\mathbf{perm}}$ and $\mathcal{T}_{\mathbf{syn}}$ because they can potentially reorder individual words in the contiguous answer, thus distributing them throughout the transformed sentence and making the question unanswerable. On the other hand, $\mathcal{T}_{\mathbf{inv}}$ and $\mathcal{T}_{\mathbf{trans}}$ do not have this issue because they maintain the spans.

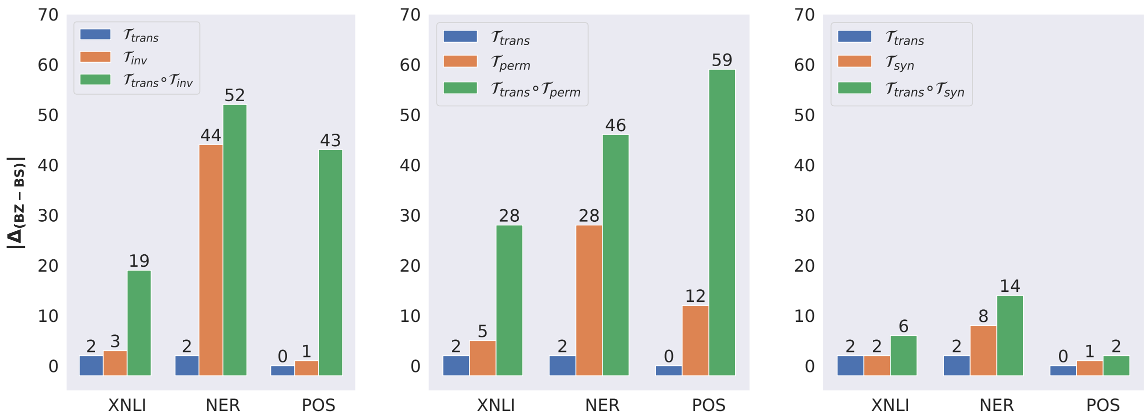


Figure 2: $|\Delta_{(\mathrm{BZ-BS})}|$ for composed transformations (§ 4.2) applied on EN as the *original* language. Larger scores imply worse zero-shot transfer. $\mathcal{T}_{\mathbf{trans}}$ = *Transliteration*, $\mathcal{T}_{\mathbf{inv}}$ = *Inversion*, $\mathcal{T}_{\mathbf{perm}}$ = *Permutation*, and $\mathcal{T}_{\mathbf{syn}}$ = *Syntax*. Sub-word overlap between the *original* and *derived* language is 0% when composed transformations are used (e.g. $\mathcal{T}_{\mathbf{trans}} \circ \mathcal{T}_{\mathbf{inv}}$) and 100% when the second constituent is used (here, $\mathcal{T}_{\mathbf{inv}}$). We observe that the composed transformations (green bars) do significantly worse than their constituents (blue and orange). $\mathcal{T}_{\mathbf{trans}} \circ \mathcal{T}_{\mathbf{inv}}$ is worse than $\mathcal{T}_{\mathbf{inv}}$ by over 16 points on XNLI and 42 points on POS, with similar trends for $\mathcal{T}_{\mathbf{trans}} \circ \mathcal{T}_{\mathbf{perm}}$. $\mathcal{T}_{\mathbf{trans}} \circ \mathcal{T}_{\mathbf{syn}}$ doesn't suffer as much, but its performance degradation when compared to *Syntax* is still large (ranges between 1 point on POS to 6 points on NER). **These results show that the absence of sub-word overlap can significantly hurt performance when languages differ in their word orders.**

4.1 Sub-word overlap is not strictly necessary for strong zero-shot transfer

Sub-word overlap is the number of common tokens between two different language corpora. If $\mathcal{E}_1$ and $\mathcal{E}_2$ are sets of tokens which appear in the two corpora, then: Sub-word overlap $= |\mathcal{E}_1 \cap \mathcal{E}_2| / |\mathcal{E}_1 \cup \mathcal{E}_2|$ (Pires et al., 2019). The *Transliteration* transformation ($\mathcal{T}_{\mathbf{trans}}$) creates *original-derived* language pairs that have 0% sub-word overlap (equivalently, different scripts), but follow the same word order.

Table 3 displays $\Delta_{(\mathrm{BZ-BS})}$ scores for $\mathcal{T}_{\mathbf{trans}}$, averaged over four languages (Appendix B contains a breakdown). We observe that $\Delta_{(\mathrm{BZ-BS})} \approx 0$ for all tasks while $\Delta_{(\mathrm{MZ-BS})}$ is highly negative, implying that zero-shot transfer is strong and on par with supervised learning. This result indicates that zero-shot transfer is possible even when languages with different scripts have similar word orders (in line with K et al. (2020)). However, it is unrealistic for natural languages to differ only in their script and not other properties (e.g., word order).

4.2 Absence of sub-word overlap significantly hurts zero-shot performance when languages differ in their word-orders

To simulate a more realistic scenario, we create *original* and *derived* language pairs which differ both in their scripts (0% sub-word overlap) and in word order. We achieve this by composing two transformations on the *original* language corpus, one of which is *Transliteration* ($\mathcal{T}_{\mathbf{trans}}$). We experiment with three different compositions, (a) $\mathcal{T}_{\mathbf{trans}} \circ \mathcal{T}_{\mathbf{inv}}$, (b) $\mathcal{T}_{\mathbf{trans}} \circ \mathcal{T}_{\mathbf{perm}}$, and (c) $\mathcal{T}_{\mathbf{trans}} \circ \mathcal{T}_{\mathbf{syn}}$. Here, $\boldsymbol{\alpha} \circ \boldsymbol{\beta}$ means that transformation $\boldsymbol{\beta}$ is applied before $\boldsymbol{\alpha}$. A composed transformation ($\mathcal{T}_{\mathbf{trans}} \circ \boldsymbol{\beta}$) differs from its second constituent ($\boldsymbol{\beta}$) in that the former produces a *derived* language which has 0% sub-word overlap with the *original* language whereas the latter has a 100% sub-word overlap.

Results Our results (Figure 2, breakdown in Appendix C) show that zero-shot performance is significantly hurt for composed transformations when compared to its constituents. $|\Delta_{(\mathrm{BZ-BS})}|$ is much larger for $\mathcal{T}_{\mathbf{trans}} \circ \mathcal{T}_{\mathbf{inv}}$ when compared to $\mathcal{T}_{\mathbf{trans}}$ or $\mathcal{T}_{\mathbf{inv}}$ individually. For example, for XNLI, $|\Delta_{(\mathrm{BZ-BS})}| = 19$ for the composed transformation and just 2 and 3 for $\mathcal{T}_{\mathbf{trans}}$ and $\mathcal{T}_{\mathbf{inv}}$ individually. $\mathcal{T}_{\mathbf{trans}} \circ \mathcal{T}_{\mathbf{perm}}$ is worse by ≈ 20 points on XNLI and NER, and over 40 points on POS when compared to $\mathcal{T}_{\mathbf{perm}}$. $\mathcal{T}_{\mathbf{trans}} \circ \mathcal{T}_{\mathbf{syn}}$ suffers lesser than the other two composed transformations, but it is still worse than $\mathcal{T}_{\mathbf{syn}}$ by 3, 6, and 1 point on XNLI, NER, and POS. In conclusion, the absence of sub-word overlap significantly degrades zero-shot performance in the realistic case of languages with different word orders.

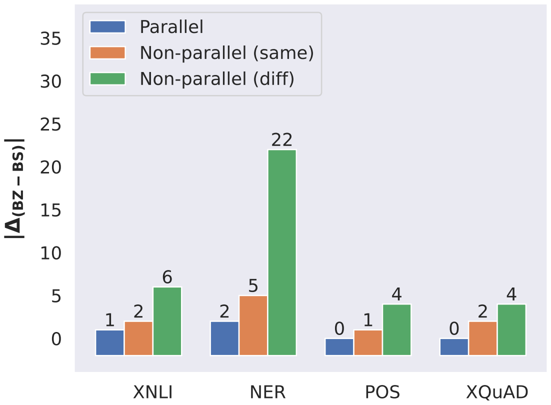


Figure 3: $|\Delta_{(\text{BZ}-\text{BS})}|$ for $\mathcal{T}_{\textbf{trans}}$ under different conditions on the source of *original* and *derived* language pre-training corpora (hereon, corpora) (§ 4.3), averaged over four languages. Larger values imply worse zero-shot transfer. The breakdown of scores for different languages is in Appendix D. (1) *Non-parallel (diff)* (green bar), which uses corpora from different domains is worse than (2) *Non-parallel (same)* (orange bar), which uses different sets of sentences sampled from the same domain, which is in turn worse than (3) *Parallel*, which uses the same sentences. Having pre-training corpora from the same domain like Wikipedia (*Non-parallel (same)*) gives performance boosts between 2 points for QA to 17 points for NER when compared to *Non-parallel (diff)*.

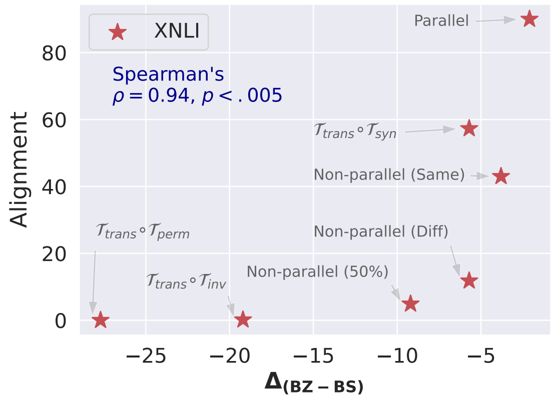


Figure 4: $\Delta_{(\text{BZ}-\text{BS})}$ for *Transliteration* ($\mathcal{T}_{\textbf{trans}}$) variants on XNLI. Larger values (less negative) imply better zero-shot transfer. We see that alignment (§ 4.4) between token embeddings of different languages is correlated with $\Delta_{(\text{BZ}-\text{BS})}$, and hence with better zero-shot transfer. For example, $\mathcal{T}_{\textbf{trans}} \circ \mathcal{T}_{\textbf{inv}}$ (bottom left) which has poor zero-shot transfer also has lower alignment, whereas *Parallel* (top right) which has strong transfer is accompanied with higher alignment. We find a strong and statistically significant Spearman's correlation of $\rho_s = 0.94, p < .005$ on XNLI, $\rho_s = 0.93, p < .005$ on NER, and $\rho_s = 0.89, p < .01$ on POS. Plots for other tasks are in Appendix E.

4.3 Data from the same domain boosts bilingual performance

Previously, we considered transformations ($\mathcal{T}$) that modified the *original* pre-training corpus to get a parallel corpus, $\mathcal{C}_{\textbf{deriv}} = \mathcal{T}(\mathcal{C}_{\textbf{orig}})$, such that there is a one-to-one correspondence between sentences in $\mathcal{C}_{\textbf{orig}}$ and $\mathcal{C}_{\textbf{deriv}}$ (we call this setting *parallel*). Since procuring large parallel corpora is expensive in practice, we consider two other settings which use different corpora for *original* and *derived*.

Setup Consider two text corpora of the same size, $\mathcal{C}^{\textbf{1}}_{\textbf{orig}}$ and $\mathcal{C}^{\textbf{2}}_{\textbf{orig}}$. We compare two settings: (1) The *parallel* setting pre-trains a bilingual model on $\mathcal{C}^{\textbf{1}}_{\textbf{orig}} + \mathcal{T}(\mathcal{C}^{\textbf{1}}_{\textbf{orig}})$, whereas the (2) *non-parallel* corpus setting uses $\mathcal{C}^{\textbf{1}}_{\textbf{orig}} + \mathcal{T}(\mathcal{C}^{\textbf{2}}_{\textbf{orig}})$. We consider two variants of *non-parallel*, (1) *non-parallel (same)* which uses different splits of Wikipedia data (hence, *same* domain), and (2) *non-parallel (diff)* which uses Wikipedia data for the *original* and common crawl data (web text) for the *derived* language (hence, *diff*erent domain). We use the *Transliteration* transformation ($\mathcal{T}_{\textbf{trans}}$) to generate the *derived* language corpus and report $|\Delta_{(\text{BZ}-\text{BS})}|$ averaged over all languages in Figure 3.

Results We observe consistently on all tasks that the *parallel* setting (blue bar) performs better than both the non-parallel settings. *Non-parallel (same)* performs better than *non-parallel (diff)*, with gains ranging between 2 points on XQuAD to 17 points on NER. This result shows that even for *original* and *derived* language pairs which differ only in their script, having parallel pre-training corpora leads to the best zero-shot transfer. Since large-scale parallel unsupervised data is hard to procure, the best alternative is to use corpora from similar domains (Wikipedia) rather than different ones (Wikipedia v.s. web text).

4.4 Zero-shot performance is strongly correlated with embedding alignment

Our previous results (§ 4.2, 4.3) showed cases where zero-shot transfer between languages is poor when there is no sub-word overlap. To investigate this further, we analyze the static word embeddings learned by bilingual models and find that zero-shot transfer between languages is strongly correlated with the alignment between word embeddings for the *original* and *derived* languages.

Setup The *original* and the *derived* languages have a one-to-one correspondence between their sub-word vocabularies when we use *transliteration* ($\mathcal{T}_{\textbf{trans}}$). For a token embedding in the *original*-language embedding matrix, its alignment score is 100% if it retrieves the corresponding token embedding in the *derived* language when a nearest-neighbor search is performed, and 0% otherwise. We average the alignment score over all the tokens and call it *alignment*.

Results We measure the *alignment* of bilingual models pre-trained on different *original-derived* language pairs created using *transliteration*, namely the composed transformations (§ 4.2), *parallel*, and *non-parallel* (§ 4.3). We plot the *alignment* along with the corresponding $\Delta_{(\text{BZ}-\text{BS})}$ scores for XNLI in Figure 4. Results for other tasks are in Appendix E.

We observe that higher *alignment* is associated with lower $\Delta_{(\text{BZ}-\text{BS})}$, implying better zero-shot transfer. *Alignment* is lower for composed transformations like $\mathcal{T}_{\textbf{trans}} \circ \mathcal{T}_{\textbf{inv}}$ and $\mathcal{T}_{\textbf{trans}} \circ \mathcal{T}_{\textbf{perm}}$ which have large and negative $\Delta_{(\text{BZ}-\text{BS})}$. *Alignment* also explains the results in Section 4.3, with *non-parallel* variants having lower alignment scores than *parallel*, which is in line with their lower $\Delta_{(\text{BZ}-\text{BS})}$. Overall, we find a strong and significant Spearman's rank correlation between *alignment* and $\Delta_{(\text{BZ}-\text{BS})}$, with $\rho = 0.94, p < .005$ for XNLI, $\rho = 0.93, p < .005$ for NER, and $\rho = 0.89, p < .01$ for POS, indicating that increasing the embedding alignment between languages helps improve zero-shot transfer.

4.5 Connections to results on natural language pairs

Effect of sub-word overlap In § 4.2, we showed that when languages have different scripts (0% sub-word overlap), zero-shot transfer significantly degrades when they additionally have different word orders. However, the zero-shot transfer is good when languages differ only in the script and have similar or the same word order. This is in line with anecdotal evidence in Pires et al. (2019), where zero-shot transfer works well between *English* and *Bulgarian* (different script but same subject-verb-object order – SVO), but is poor between *English* and *Japanese* (different script *and* word order – SVO v.s. SOV). Our result also corroborates findings in Conneau et al. (2020b) that artificially increasing sub-word overlap between natural languages (which have different word orders) improves performance (e.g., 3 points on XNLI).

Effect of token embedding alignment In § 4.4, we showed that zero-shot transfer is strongly correlated with word embedding alignment between languages. This explains the usefulness of recent studies which try to improve multilingual pre-training with the help of auxiliary objectives, which improve word or sentence embedding alignment.

DICT-MLM (Chaudhary et al., 2020) and RelateLM (Khemchandani et al., 2021) require the model to predict cross-lingual synonyms as an auxiliary objective, thus indirectly improving word-embedding alignment and the zero-shot performance on multiple tasks. Hu et al. (2021) add an auxiliary objective that implicitly improves word embedding alignment and show that they can achieve performance similar to larger models. Cao et al. (2019) explicitly improve contextual word embedding alignment with the help of word-level alignment information in machine-translated cross-lingual sentence pairs. Since they apply this post hoc and not during pre-training, the improvement, albeit significant, is small (2 points on XNLI). While these studies do not fully utilize word and sentence embedding alignment information, our results lead us to posit that they are a step in the right direction and that baking alignment information more explicitly into pre-training will be beneficial.

5 Conclusion

Through a systematic study of zero-shot transfer between four diverse natural languages and their counterparts created by modifying specific properties like the script, word order, and syntax, we showed that (1) absence of sub-word overlap hurts zero-shot performance when languages differ in their word order, and (2) zero-shot performance is strongly correlated with word embedding alignment between languages. Some recent studies have implicitly or unknowingly attempted to improve alignment and have shown slight improvements in zero-shot transfer performance. However, our results lead us to posit that explicitly improving word embedding alignment during pre-training by using either supervised (e.g., parallel sentences and translation dictionaries) or unsupervised data will significantly improve zero-shot transfer.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020a. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4623–4637. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020b. On the cross-lingual transferability of monolingual representations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Samuel Bowman, Christopher Potts, and Christopher D Manning. 2015. Recursive neural networks can learn logical semantics. In *Proceedings of the 3rd workshop on continuous vector space models and their compositionality*, pages 12–21.

Steven Cao, Nikita Kitaev, and Dan Klein. 2019. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Aditi Chaudhary, Karthik Raman, Krishna Srinivasan, and Jiecao Chen. 2020. DICT-MLM: improved multilingual pre-training using bilingual dictionaries. *CoRR*, abs/2010.12566.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

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.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034.

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.

Sumanth Doddapaneni, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra. 2021. A primer on pretrained multilingual language models. *arXiv preprint arXiv:2107.00676*.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for bert's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437.

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

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.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: an empirical study. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. inlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4948–4961.

Yash Khemchandani, Sarvesh Mehtani, Vaidehi Patil, Abhijeet Awasthi, Partha Talukdar, and Sunita Sarawagi. 2021. Exploiting language relatedness for low web-resource language model adaptation: An Indic languages study. 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 1312–1323, Online. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting* *of the Association for Computational Linguistics*, pages 3125–3135.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, and Mohammed Attia et al. 2018. Universal dependencies 2.2. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Dingquan Wang and Jason Eisner. 2016. The galactic dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, 4:491–505.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 833–844. Association for Computational Linguistics.

Appendices

A Mathematical Specification for Transformation of Downstream Datasets

Text classification Text classification tasks like news classification or sentiment analysis typically have instances which contain a single sentence and a label. Instances in other classification tasks like natural language inference (NLI) (Bowman et al., 2015) contain two sentences and one label. For such tasks, we apply the transformation ($\mathcal{T}$) on each sentence within every instance, and leave the annotated label as is. Therefore, for a dataset of size n which contains m sentences per instance, we have:

$$\mathcal{D}_{\textbf{orig}} = \{(s_{i1}, \ldots, s_{im}, y_i) \mid i = 1 : N\}$$
$$\mathcal{D}_{\textbf{deriv}} = \{(\mathcal{T}(s_{i1}), \ldots, \mathcal{T}(s_{im}), y_i) \mid i = 1 : N\}$$

Token-classification tasks Tasks like named-entity recognition (NER) and part-of-speech tagging (POS tagging) have labels associated with *each* token in the sentence. For these datasets, we ensure that any transformation ($\mathcal{T}$) that changes the order of the tokens also changes the order of the corresponding labels.

We define a few quantities to express the transformation mathematically. Let $s_i = (w_{i1}, \ldots, w_{ik})$ be a sentence comprised of k tokens and $y_i = (y_{i1}, \ldots, y_{ik})$ be labels corresponding to the tokens in the sentence. We define a new transformation ($\mathcal{T}_{aug}$) which operates on the label augmented sentence, $s_i^{aug} = ((w_{i1}, y_{i1}), \ldots, (w_{ik}, y_{ik}))$. Let $s_i^{aug}[j]$ correspond to the j^{th} element in the sequence, and $s_i^{aug}[j][\texttt{word}]$ and $s_i^{aug}[j][\texttt{label}]$ correspond to the word and label of the j^{th} element. Let $\mathcal{T}_{aug}(s_i^{aug})[j][\texttt{orig}]$ denote the index of the j^{th} element in the transformed sequence with respect to the original sequence s_i^{aug}. Then, the new transformation $\mathcal{T}_{aug}$ is such that,

$$\mathcal{T}_{aug}(s_i^{aug})[j][\texttt{orig}] = \mathcal{T}(s_i)[j][\texttt{orig}]$$
$$\text{Let } \texttt{orig\_j} = \mathcal{T}_{aug}(s_i^{aug})[j][\texttt{orig}]$$
$$\mathcal{T}_{aug}(s_i^{aug})[j][\texttt{label}] = s_i^{aug}[\texttt{orig\_j}][\texttt{label}]$$

We transform the dataset using $\mathcal{T}_{aug}$:

$$\mathcal{D}_{\textbf{orig}} = \{s_i^{aug} \mid i = 1 : N\}$$
$$\mathcal{D}_{\textbf{deriv}} = \{\mathcal{T}_{aug}(s_i^{aug}) \mid i = 1 : N\}$$

B Zero-shot transfer results for different transformations

Table 5 in the appendix is the extended version of Table 3 in the main paper with a breakdown for all languages. It reports $\Delta_{(\mathrm{BZ-BS})}$, $\Delta_{(\mathrm{MZ-BS})}$, and *BZ* for different languages and transformations considered.

C Composed Transformations

Table 6 in the appendix presents the breakdown of results in Figure 2 of the main paper. It reports $\Delta_{(\mathrm{BZ-BS})}$ scores for composed transformations and their constituents.

D Comparing different sources for *original* and *derived* language corpora

Table 8 in the appendix contains the breakdown of results in Figure 3 of the main paper. It reports $\Delta_{(\mathrm{BZ-BS})}$ for different languages on different tasks for the settings mentioned in Section 4.3.

E Alignment Correlation

We present alignment results (Section 4.4) for all XNLI, NER, and POS in Figure 5. We observe strong correlations between alignment and zero-shot transfer, with $\rho_s = 0.94, p < .005$ on XNLI, $\rho_s = 0.93, p < .005$ on NER, and $\rho_s = 0.89, p < .01$ on POS. We present the raw scores in Table 7.

F Hyperparameters for XTREME

- XNLI: Learning rate – `2e-5`, maximum sequence length – 128, epochs – 5, batch size – 32.
- NER: Learning rate – `2e-5`, maximum sequence length – 128, epochs – 10, batch size – 32.
- POS: Learning rate – `2e-5`, maximum sequence length – 128, epochs – 10, batch size – 32.
- Tatoeba: Maximum sequence length – 128, pooling strategy – representations from the middle layer $\left(\frac{n}{2}\right)$ of the model.
- XQuAD: Learning rate – `3e-5`, maximum sequence length – 384, epochs – 2, document stride – 128, warmup steps – 500, batch size – 16, weight decay – 0.0001.

Task	Language	Inversion			Permutation			Syntax			Transliteration		
		BZ	$\Delta_{(\text{BZ}-\text{BS})}$	$\Delta_{(\text{MZ}-\text{BS})}$	*BZ*	$\Delta_{(\text{BZ}-\text{BS})}$	$\Delta_{(\text{MZ}-\text{BS})}$	*BZ*	$\Delta_{(\text{BZ}-\text{BS})}$	$\Delta_{(\text{MZ}-\text{BS})}$	*BZ*	$\Delta_{(\text{BZ}-\text{BS})}$	$\Delta_{(\text{MZ}-\text{BS})}$
XNLI	**English**	73.2	-3.4	-14.9	68.6	-5	-7.7	74.1	-1.8	-1.5	74.1	-1.7	-42.5
	French	62.5	-9.5	-8.8	68.4	-1	-7.6	69.6	-2.2	-1.4	71.6	-1.6	-39.9
	Hindi	43.9	-15.7	-15.8	51.2	-6.2	-13.1	61.6	-0.3	-1.6	63.4	-0.1	-29.4
	Arabic	54	-12.3	-12.5	62.1	-2.3	-6	65.9	0.7	0.3	68	-0.4	-35.1
	Avg.	58.4	**-10.2**	**-13**	62.6	**-3.6**	**-8.6**	67.8	**-0.9**	**-1.1**	69.3	**-1.0**	**-36.7**
NER	**English**	39.8	-44.5	-35.9	40.2	-28.5	-33.2	61.1	-7.8	-10.3	78	-2.1	-70.2
	French	54.5	-34.4	-51.3	44.4	-36.0	-39.8	59.6	-21.9	-25.9	84.3	-3.1	-87.4
	Hindi	19.4	-63.9	-63.2	38.5	-21.9	-37.4	64.8	-8.4	-7.3	84.4	-0.5	-82.9
	Arabic	37.8	-53.6	-36.3	66.2	-18.8	-31.1	66.1	-20.1	-23	88	-1.9	-89.9
	Avg.	37.9	**-49.1**	**-46.7**	47.3	**-26.3**	**-35.4**	62.9	**-14.6**	**-16.6**	83.7	**-1.9**	**-82.6**
POS	**English**	94.4	-0.7	-24.3	78.3	-11.9	-17.6	92.9	-0.9	-2.2	94.6	-0.5	-95.1
	French	74.3	-22.7	-22.9	82	-12.2	-20.9	93.5	-3.2	-5.2	97.2	-0.2	-97.4
	Hindi	19	-74.5	-74.5	51	-14	-41.8	91.6	-3.3	-11.3	96.5	-0.1	-96.6
	Arabic	69.2	-23	-23	83.1	-6.5	-20.6	79.4	-10	-11.5	93.2	-0.8	-90.9
	Avg.	64.2	**-30.2**	**-36.2**	73.6	**-11.2**	**-25.2**	89.4	**-4.4**	**-7.6**	95.4	**-0.4**	**-95.0**
XQuAD	**English**	30.4	-43.2	-35.5	-	-	-	-	-	-	72.4	-4	-73
	French	25.2	-29.5	-29.6	-	-	-	-	-	-	60.9	-1	-55.5
	Hindi	14.5	-27.3	-27.3	-	-	-	-	-	-	57.3	10.6	-43.5
	Arabic	21	-31.2	-31.4	-	-	-	-	-	-	54	-0.5	-51.7
	Avg.	22.8	**-32.8**	**-31.0**							61.2	**1.3**	**-55.9**

Table 5: This table is an extended version of Table 3 in the main paper. Smaller (more negative) $\Delta_{(\text{BZ}-\text{BS})}$ implies worse bilingual zero-shot transfer, whereas $\Delta_{(\text{BZ}-\text{BS})} \approx 0$ implies strong transfer. $\Delta_{(\text{BZ}-\text{BS})} \gg \Delta_{(\text{MZ}-\text{BS})}$ implies that bilingual pre-training is extremely useful. Scores are highlighted based on their value (lower scores have a higher intensity of red). **(1) Discussing $\Delta_{(\text{BZ}-\text{BS})}$:** $\mathcal{T}_{\text{trans}}$ exhibits strong transfer on all tasks and languages (high $\Delta_{(\text{BZ}-\text{BS})}$ scores), and bilingual pre-training is extremely useful ($\Delta_{(\text{BZ}-\text{BS})} \gg \Delta_{(\text{MZ}-\text{BS})}$), implying that zero-shot transfer is possible between languages with different scripts but the same word order. $\mathcal{T}_{\text{inv}}$ and $\mathcal{T}_{\text{perm}}$ suffer on all tasks (small $\Delta_{(\text{BZ}-\text{BS})}$ scores) whereas $\mathcal{T}_{\text{syn}}$ suffers significantly lesser, which provides evidence that local changes to the word order made by *Syntax* ($\mathcal{T}_{\text{syn}}$) hurts zero-shot transfer significantly lesser than global changes made by *Inversion* ($\mathcal{T}_{\text{inv}}$) and *Permutation* ($\mathcal{T}_{\text{perm}}$). **(1) Discussing $\Delta_{(\text{MZ}-\text{BS})}$:** $\Delta_{(\text{BZ}-\text{BS})}$ is much larger than $\Delta_{(\text{MZ}-\text{BS})}$ for $\mathcal{T}_{\text{trans}}$, implying that bilingual pre-training (hereon, pre-training) is extremely useful. $\Delta_{(\text{BZ}-\text{BS})}$ and $\Delta_{(\text{MZ}-\text{BS})}$ are similar for $\mathcal{T}_{\text{inv}}$ and $\mathcal{T}_{\text{syn}}$, implying that pre-training is not beneficial for these transformations. $\Delta_{(\text{BZ}-\text{BS})}$ is slightly larger than $\Delta_{(\text{MZ}-\text{BS})}$ for $\mathcal{T}_{\text{perm}}$, which means that pre-training is moderately useful.

$\mathcal{T}$	XNLI		NER		POS	
	BZ	$\Delta_{(\text{BZ}-\text{BS})}$	*BZ*	$\Delta_{(\text{BZ}-\text{BS})}$	*BZ*	$\Delta_{(\text{BZ}-\text{BS})}$
$\mathcal{T}_{\text{trans}}$	74.1	-2.1	78	-2.3	94.6	-0.5
$\mathcal{T}_{\text{inv}}$	73.2	-3.4	39.8	-44.5	94.4	-0.7
$\mathcal{T}_{\text{trans}} \circ \mathcal{T}_{\text{inv}}$	55.7	-19.2	32.5	-51.5	52.2	-42.7
$\mathcal{T}_{\text{perm}}$	68.6	-5	40.2	-28.5	78.3	-11.9
$\mathcal{T}_{\text{trans}} \circ \mathcal{T}_{\text{perm}}$	44	-27.7	17.1	-46.3	29.5	-59
$\mathcal{T}_{\text{syn}}$	74.1	-1.8	61.1	-7.8	92.9	-0.9
$\mathcal{T}_{\text{trans}} \circ \mathcal{T}_{\text{syn}}$	69.8	-5.7	53.5	-14.2	91.5	-2

Table 6: Breakdown of results in Figure 2 of the main paper. *BZ* is the zero-shot performance. $\Delta_{(\text{BZ}-\text{BS})}$, $\Delta_{(\text{MZ}-\text{BS})}$, and *BZ* are described in Section 3.3 and Table 2. Composing transformations always hurts $\Delta_{(\text{BZ}-\text{BS})}$ when compared to individual transformations.

Transliteration Variant	$\Delta_{(\text{BZ}-\text{BS})}$ (↑) XNLI	NER	POS	Alignment (↑)
Parallel	-2.1	-2.3	-0.5	90.0
Trans ∘ Syntax	-5.7	-14.2	-2	57.3
Non-parallel (Same)	-3.8	-4.1	-0.7	43.0
Non-parallel (Diff)	-5.7	-14.3	-1.5	11.8
Trans ∘ Inv	-19.2	-51.5	-42.7	0.16
Trans ∘ Perm	-27.7	-46.3	-59	0.01

Table 7: $\Delta_{(\text{BZ}-\text{BS})}$ and *alignment* scores for different *Transliteration* variants. The table contains raw scores for results in Section 4.4 of the main paper. Rows are sorted in descending order based on *alignment*. We observe strong correlations between alignment and zero-shot transfer, with $\rho_s = 0.94, p < .005$ on XNLI, $\rho_s = 0.93, p < .005$ on NER, and $\rho_s = 0.89, p < .01$ on POS.

Task	Language	XNLI	NER	POS	XQuAD
		$\Delta_{(\text{BZ}-\text{BS})}$	$\Delta_{(\text{BZ}-\text{BS})}$	$\Delta_{(\text{BZ}-\text{BS})}$	$\Delta_{(\text{BZ}-\text{BS})}$
Parallel	**English**	-1.7	-2.1	-0.5	-4
	French	-1.6	-3.1	-0.2	-1
	Hindi	-0.1	-0.5	-0.1	10.6
	Arabic	-0.4	-1.9	-0.8	-0.5
	Avg.	**-1.0**	**-1.9**	**-0.4**	**1.3**
Non-parallel (Same)	**English**	-3.8	-4.1	-0.7	-6.9
	French	-1	-6.3	-0.5	-0.9
	Hindi	-0.4	-3.1	-0.2	4.5
	Arabic	-2	-6.1	-1.5	0.7
	Avg.	**-1.8**	**-4.9**	**-0.7**	**-0.6**
Non-parallel (Diff)	**English**	-5.7	-14.3	-1.5	-9.3
	French	-10.9	-30.3	-10.5	-5.2
	Hindi	-0.5	-8.6	-1	5
	Arabic	-6.3	-34.7	-3.7	-1.9
	Avg.	**-5.9**	**-22.0**	**-4.2**	**-2.9**

Table 8: $|\Delta_{(\text{BZ}-\text{BS})}|$ for $\mathcal{T}_{\textbf{trans}}$ under different conditions on the source of *original* and *derived* language pre-training corpora (§ 4.3). Larger values imply worse zero-shot transfer. For all languages: (1) *Non-parallel (diff)*, which uses corpora from different domains is worse than (2) *Non-parallel (same)*, which uses different sets of sentences sampled from the same domain, which is in turn worse than (3) *Parallel*, which uses the same sentences.

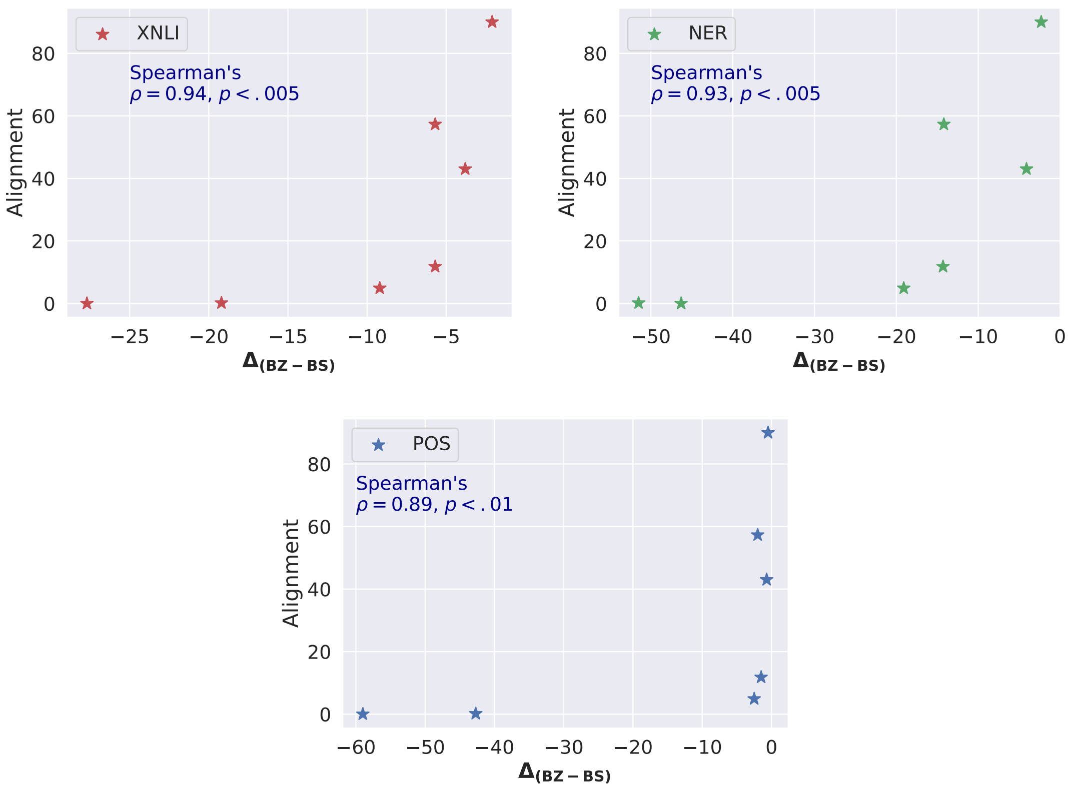


Figure 5: Alignment v.s. $\Delta_{(\text{BZ}-\text{BS})}$ plots for XNLI, NER, and POS. We observe strong correlations between alignment and zero-shot transfer, with $\rho_s = 0.94, p < .005$ on XNLI, $\rho_s = 0.93, p < .005$ on NER, and $\rho_s = 0.89, p < .01$ on POS.