# The Impact of Cross-Lingual Adjustment of Contextual Word Representations on Zero-Shot Transfer

Anonymous ACL submission

## Abstract

We evaluate effectiveness of an existing approach to cross-lingual adjustment of mBERT using four typologically different languages (Spanish, Russian, Vietnamese, and Hindi) and three NLP tasks (QA, NLI, and NER). The adjustment uses a *small* parallel corpus to make embeddings of related words across languages similar to each other. It improves NLI in four languages and NER in three languages, while QA performance never improves and sometimes degrades. Analysis of distances between contextualized embeddings of related and unrelated words across languages showed that fine-tuning leads to "foregetting" some of the cross-lingual alignment information, which—we conjecture—can negatively affect the effectiveness of the zero-shot transfer. Based on this observation, we further improved performance on NLI using continual learning. Our study contributes to a better understanding of cross-lingual transfer capabilities of large multilingual language models and of effectiveness of their cross-lingual adjustment in various NLP tasks.

## 1 Introduction

Large language models such as mBERT or XLM-R are pre-trained on multilingual corpora without parallel data annotation and enable zero-shot *cross-lingual* transfer (Libovický et al., 2019; Pires et al., 2019). Zero-shot transfer works even for languages not seen at the pre-trainig stage (Ebrahimi et al., 2021; Muller et al., 2021). Contextualized word representations produced by the models can be further aligned using a modest amount of parallel data, which was shown to improve zero-shot transfer for syntactic parsing, natural language inference (NLI), and NER (Kulshreshtha et al., 2020; Wang et al., 2019b,a). This approach requires less data and is a more computationally efficient alternative to training a machine translation system or a pre-training a large multilingual model on a large parallel corpus. The common approach that has been used since advent of static monolingual word embeddings is to find a rotation matrix using a bilingual dictionary or a parallel corpus that brings vector representation of related words in different languages closer to each other. Different from post hoc rotation-based alignment, Cao et al. (2020) employed parallel data for direct cross-lingual adjustment of the mBERT model. They showed it to be more effective than rotation in cross-lingual NLI and parallel sentence retrieval tasks in five European languages.

However, we are not aware of any systematic study of the effectiveness of this procedure across typologically diverse languages and different NLP tasks. To fill this gap, we first adjust mBERT using parallel data (English vs. Spanish, Russian, Vietnamese, and Hindi) with an objective to make embeddings of semantically similar words (in different languages) to be closer to each other as proposed by Cao et al. (2020). Then, we fine-tune cross-lingually adjusted mBERT models for three NLP tasks (NLI, NER, and QA) using English data in the regular and *continual-learning* mode (Ratcliff, 1990; Robins, 1995; Parisi et al., 2019) by using an auxiliary cross-lingual adjustment loss during fine-tuning (Caruana, 1996). Finally, we apply the trained models to the test data in four target languages in a zero-shot fashion (i.e., without fine-tuning in the target language).

We perform each experiment with five seeds and assess statistical significance of the difference from a baseline. In our study we ask the following research questions:

- R1 How does cross-lingually adjusted mBERT fine-tuned on English data and zero-shot transferred to a target language perform on various NLP tasks and target languages?
- R2 How does the size of the parallel corpora used for adjustment affect outcomes?

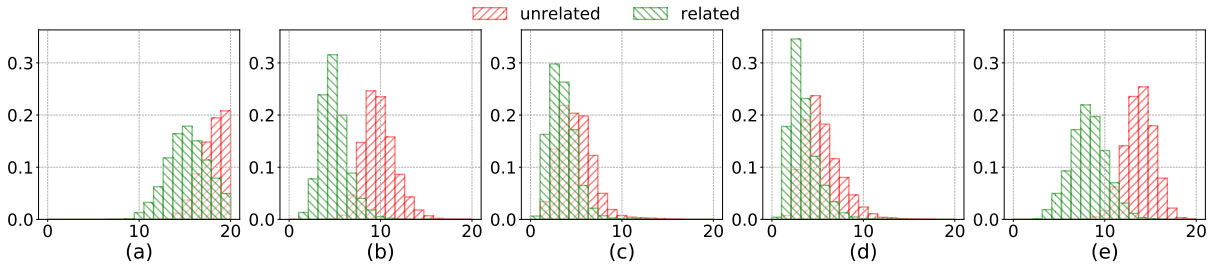


Figure 1: Histograms of $L_2$ distances between pairs of mBERT *last-layer* representations for randomly sampled related (i.e., aligned) and unrelated word pairs from WikiMatrix (Hi-En): (a) original, (b) after cross-lingual adjustment, (c) after fine-tuning on English NLI data, (d) after cross-lingual adjustment and subsequent fine-tuning on English NLI data, (e) after cross-lingual adjustment and subsequent *continual* fine-tuning on English NLI data.

R3 How does adjustment of mBERT on parallel data and fine-tuning for a specific task affect similarity of contextualized embeddings of semantically related and unrelated words across languages?

R4 Inspired by our observation (see Fig. 1c-1d) that fine-tuning draws embeddings of *both* related and unrelated words closer to each other, which may negatively affect the cross-lingual transfer, we wonder if continual learning—with an auxiliary cross-lingual adjustment loss—can improve effectiveness of the zero-shot transfer.

Our experiments demonstrated the following:

- The cross-lingual adjustment of mBERT improves NLI in four languages and NER in three languages. Yet, there is no statistically significant improvement for QA and a statistically significant deterioration on three out of eight QA datasets. Experiments with Hindi and extended BERT (Wang et al., 2020) indicate this could be due to insufficient vocabulary representation for some languages in mBERT.

- As the amount of parallel data increases, this benefits both NLI and NER, whereas QA performance peaks at roughly 5K parallel sentences and further decreases as the number of parallel sentences increases.

- When comparing $L_2$ distances between contextualized-embeddings of words across languages (Fig. 1b), we see that the cross-lingual adjustment of mBERT decreases the $L_2$ distance between related words while keeping unrelated words apart, which is in line with prior work (Zhao et al., 2021).

- However, we have found no prior work that inspected histograms obtained after fine-tuning. Quite surprisingly, we observe that fine-tuning of mBERT for a specific task draws embeddings of *both* related and unrelated words much closer to each other (Fig. 1c and Fig. 1d). Thus, fine-tuning causes the model to "forget" some of the cross-lingual information learned during adjustment.

- In that, continual learning allows the model to learn a target task while maintaining the separation of related and unrelated words (Fig. 1e). Continual learning consistently improves performance on NLI data, but we obtain no improvement on either QA or NER. In fact, we observe that improving separation between the related and unrelated words across languages—which is the object of the cross-lingual adjustment that is optionally reinforced with continual learning—does not help cross-lingual transfer among all tasks and training regimes.

In summary, our study contributes to a better understanding of (1) cross-lingual transfer capabilities of large multilingual language models and of (2) effectiveness of their cross-lingual adjustment in various NLP tasks. Inspired by our histogram analysis, we were able to improve performance on NLI data using continual learning, which is a novel finding in the context of zero-shot transfer with cross-lingual adjustment using the approach of Cao et al. (2020).

## 2 Related Work

### 2.1 Cross-Lingual Zero-Shot Transfer with Multilingual Models

The success of mBERT in cross-language zero-shot regime on various tasks has inspired many papers that attempted to explain its cross-lingual abilities and limitations (Wu and Dredze, 2019; Conneau et al., 2020; K et al., 2020; Libovický et al., 2019; Dufter and Schütze, 2020; Chi et al., 2020; Pires et al., 2019; Artetxe et al., 2020; Chi et al., 2020). These studies showed that the multilingual models learn high-level abstractions common to all languages, which make transfer possible even when languages share no vocabulary. However, the gap between performance on English and a target language is smaller if the languages are cognate, i.e. share a substantial portion of model's vocabulary, have similar syntactic structures, and are from the same language family (Wu and Dredze, 2019; Lauscher et al., 2020). Moreover, the size of target language data used for pre-training and the size of the model vocabulary allocated to the language also positively impacts cross-lingual learning performance (Lauscher et al., 2020; Artetxe et al., 2020).

Zero-shot transfer of mBERT or other multilingual transformer-based models from English to a different language was applied inter alia to POS tagging, cross-lingual information retrieval, dependency parsing, NER, NLI, and QA (Wu and Dredze, 2019; Wang et al., 2019b; Pires et al., 2019; Hsu et al., 2019; Litschko et al., 2021). XTREME data suite (Hu et al., 2020) and its successor XTREME-R (Ruder et al., 2021) are dedicated collections of tasks and corresponding datasets for evaluation of zero-shot transfer capabilities of large multilingual models from English to tens of languages. XTREME includes NLI, NER, and QA datsets used in the current study. Although transfer from English is not always an optimal choice (Lin et al., 2019; Turc et al., 2021), English still remains the most popular source language. Furthermore, despite there have been developed quite a few new models that differ in architectures, supported languages, and training data (Doddapaneni et al., 2021), mBERT remains the most popular cross-lingual model.

### 2.2 Cross-lingual Alignment of Embeddings

Mikolov et al. (2013) demonstrated that vector spaces can encode semantic relationships between words and that there are similarities in the geometry of these vectors spaces across languages. A variety of approaches have been proposed for aligning monolingual representations based on bilingual dictionaries and parallel sentences. The most widely used approach—which requires only a bilingual dictionary—consists in finding a rotation matrix that aligns vectors of two monolingual models (Mikolov et al., 2013). Lample et al. (2018) proposed an alignment method based on adversarial training, which does not require parallel data. A comprehensive overview of alignment methods for pre-Transformer models can be found in (Ruder et al., 2019).

Schuster et al. (2019) applied rotation to align contextualized ELMo embeddings (Peters et al., 2018) using "anchors" (averaged vectors of tokens in different contexts) and bilingual dictionaries. They showed improved results of cross-lingual dependency parsing using English as source and several European languages as target languages. Wang et al. (2019a) aligned English BERT and mBERT representations using rotation method and Europarl parallel data (Koehn, 2005). They employed the resulting embeddings in a cross-lingual dependency parsing model. The parser with aligned embeddings consistently outperformed zero-shot mBERT on 15 out of 17 target languages. Instead of aligning on a word level, Aldarmaki and Diab (2019) performed a sentence-level alignment of ELMo embeddings and evaluated this approach on the parallel sentence retrieval task.

Cao et al. (2020) proposed to directly modify the mBERT model by bringing the vectors of semantically related words in different languages closer to each other. This was motivated by the observation that embedding spaces of different languages are not always isometric (Søgaard et al., 2018) and, hence, are not always amenable to alignment via rotation. The authors showed that mBERT simultaneously adjusted on five European languages consistently outperformed other alignment approaches on XNLI data. In the current study, we implement the approach with some modifications.

Liu et al. (2021) showed that combining continual learning with fine-tuning improved zero-shot transfer performance for NER and POS tagging. In that, they used a cross-lingual sentence retrieval (XSR) and/or masked-language model (MLM) task as additional tasks. Although XSR can be seen as an alternative to the cross-lingual adjustment of

Cao et al. (2020), the authors did not evaluate the effectiveness of zero-shot transfer after adjusting the model with XSR. In contrast, we evaluate the marginal effectiveness of continual learning with respect to already cross-lingually adjusted mBERT.

Kulshreshtha et al. (2020) compared different alignment methods (rotation vs. adjustment) on NER and slot filling tasks. According to their results, rotation-based alignment performs better on the NER task, while model adjustment performs better on slot filling. Zhao et al. (2021) continued this line of research and proposed several improvements of the model adjustment method: 1) z-normalization of vectors and 2) text normalization to make the input more structurally 'similar' to English training data. Experiments on XNLI dataset and translated sentence retrieval showed that vector normalization leads to more consistent improvements over zero-shot baseline compared to text normalization. Faisal and Anastasopoulos (2021) applied cross-lingually adjusted mBERT and XLM-R to cross-lingual open-domain QA and obtained improvements both on paragraph and span selection subtasks. However, they trained their models on machine-translated data, which is different from our zero-shot settings.

## 3 Methods

In this study, we use a multilingual BERT (mBERT) as the main model (Devlin et al., 2019). mBERT is a case-sensitive "base" 12-layer Transformer model (Vaswani et al., 2017) with 178M parameters.[1] It was trained with a masked language model objective on 104 languages with a shared WordPiece (Wu et al., 2016) vocabulary (using 104 Wikipedias). To balance the distribution of languages, high-resource languages were undersampled and low-resource languages were oversampled.[2] For a number of NLP tasks, cross-lingual transfer of mBERT can be competitive with training a monolingual model using the training data in the target language. [3]

We align cross-lingual embeddings by directly modifying/adjusting the language model itself, following the approach by Cao et al. (2020). The approach—which differs from finding a rotation matrix—proved to be effective in the XNLI task. However, there are some differences in our implementation. In all cases, we work with one pair of languages at a time while Cao et al. (2020) adjusted mBERT for five languages at once. Our approach allows us to carry out a parameter-sensitivity analysis individually for each of the target languages.

BERT uses WordPiece tokenization (Wu et al., 2016), which splits sufficiently long words into *subword* tokens. We first word-align parallel data with *fast_align* (Dyer et al., 2013) and then average all subword tokens' vectors.[4]

Based on alignments in parallel data, we obtain a collection of word pairs $(s_i, t_i)$: $s_i$ from the source language, $t_i$ from the target one. From these alignments we can obtain their mBERT vector representations $\mathbf{f}(s_i)$ and $\mathbf{f}(t_i)$. Then, we *adjust* the mBERT model on aligned pairs' vectors using the following loss function:

$$L = \sum_{(s_i,t_i)} \|\mathbf{f}(s_i) - \mathbf{f}(t_i)\|_2^2 + \sum_{s_j} \|\mathbf{f}(s_j) - \mathbf{f}^0(s_j)\|_2^2, \quad (1)$$

where the first term "pulls" the embeddings in the source and target language together, while the second (regularization) term prevents source (English) representations from deviating far from their initial values in the 'original' mBERT $\mathbf{f}^0$. Finally, the cross-lingually adjusted mBERT model is fine-tuned for a specific task.

Training neural networks via empirical loss minimization is known to suffer from the "catastrophic forgetting" (McCloskey and Cohen, 1989). From inspecting the histogram of $L_2$ distances between embeddings of related and unrelated words in pairs of languages (see Fig. 1 and the discussion in § 5.4), we learn that this is, indeed, the case. Specifically, fine-tuning on a target task—in contrast to the cross-lingual adjustment objective—reduces the separation between related and unrelated words. To counter this effect, we ran an additional experiment in a continual-learning mode (Riabi et al., 2021), which relies on experience replay (Ratcliff, 1990; Robins, 1995).

Technically, this entails a multi-task training (Caruana, 1996) with a combined loss function:

[1] https://huggingface.co/bert-base-multilingual-cased

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

[3] Along with original mBERT we also experimented with mBERT variants with expanded vocabulary for Hindi (Wang et al., 2020), see Appendix A.

[4] We also experimented with other options reported in the literature – fist/last tokens' vectors, as well as aligning subword tokens produced by BERT. Although these choices induced some variations in results, there is no single pattern across all tasks and languages, see § A.1.

| Lang | Family | Script | Word order | Number of Wiki pages |
|---|---|---|---|---|
| en | IE/Germanic | Latin | SVO | 6.3M |
| es | IE/Romance | Latin | SVO | 1.7M |
| ru | IE/Slavic | Cyrillic | SVO | 1.7M |
| vi | Austroasiatic | Latin | SVO | 1.3M |
| hi | IE/Indo-Aryan | Devanagari | SOV | 150K |

IE : Indo-European; Prevalent word order: SVO – subject-verb-object, SOV – subject-object-verb;

Table 1: Language information.

$$L = L_{target} + \alpha L_{align}, \quad (2)$$

where $L_{target}$ is the loss-function for the target task, e.g., NLI, $L_{align}$ is a cross-lingual loss function given by Eq. 1, and $\alpha > 0$ is a small weight. During training, we iterate over the complete (reshuffled) dataset for the target task: After computing $L_{target}$ for a current batch we randomly sample a small batch of aligned pairs of words $\{(s_i, t_i)\}$ from the parallel corpus and compute $L_{align}$.

# 4 Tasks and Data

## 4.1 Languages and Parallel Data

In our experiments we transfer models trained on English to four languages: Spanish, Russian, Vietnamese, and Hindi. This set represents four different families (including one non-Indo-European language), three scripts, and two different prevalent word orders (see Table 1). All the languages are among languages that were used to train mBERT.[5]

We use a parallel corpus (i.e., a bitext) WikiMatrix (Schwenk et al., 2021) to align embeddings. WikiMatrix is a large collection of aligned sentences in 1,620 different language pairs mined from Wikipedia. The dataset is distributed under CC-BY-SA license.

## 4.2 Natural Language Inference

Natural language inference (NLI) is a task of determining the relation between two ordered sentences (hypothesis and premise) and classifying them into: entailment, contradiction, or "no relation". English MultiNLI collection consists of 433K multi-genre sentence pairs (Williams et al., 2018). The XNLI dataset—distributed under the CC BY-NC license—complements the MultiNLI training set with newly collected 2.5K development and 5K test English examples (Conneau et al., 2018). They were professionally translated into 15 languages, including all four target languages of the current study. Additionally, for each of the target language test set, we created a new mixed-language XNLI set by randomly picking either a hypothesis or a premise and replacing it with the original English sentence. Performance on XNLI datasets is evaluated using classification *accuracy*.

[5]However, Hindi Wikipedia is an order of magnitude smaller compared to other Wikipedias, which may have led to somewhat inferior contextualized embeddings.

## 4.3 Named Entity Recognition

Named entity recognition (NER) is a task of locating named entities in unstructured text and classifying them into predefined categories such as persons, organizations, locations, etc. In our experiments, we employ the Wikiann NER corpus (Rahimi et al., 2019) that is derived from a larger "silver-standard" collection that was created fully automatically (Pan et al., 2017). Wikiann NER has data for 41 language, including all languages in the current study. The dataset is distributed under the Apache-2.0 license. The named entity types include location (LOC), person (PER), and organization (ORG). The English training set contains 20K sentences. Test sets for Spanish, Vietnamese, and Russian have 10K sentences each; for Hindi – 1K sentences. Performance is evaluated using the token-level micro-averaged F1.

## 4.4 Question Answering

Machine reading comprehension (MRC) is a variant of QA task. Given a question and a text paragraph, the system needs to return a continuous span of paragraph tokens as an answer. The first large-scale MRC dataset is the English Wikipedia-based dataset SQuAD (Rajpurkar et al., 2016), which contains about 100K paragraph-question-answer triples. SQuAD has become a *de facto* standard and inspired creation of analogous resources in other languages (Rogers et al., 2021). SQuAD is available under the CC BY-SA license. We use SQuAD as the source dataset to train MRC models. To test the models, we use XQuAD, MLQA, and TyDi QA datasets. XQuAD (Artetxe et al., 2020) is a professional translation of 240 SQuAD paragraphs and 1,190 questions-answer pairs into 10 languages (including four languages of our study). MLQA (Lewis et al., 2020) data is available for six languages including Spanish, Vietnamese, and Hindi (but it does not have Russian). There are about 5K questions for each of our languages. TyDi QA (Clark et al., 2020) includes 11 typologically

diverse languages of which we use only Russian (812 test items). SQuAD, XQuAD, and MLQA are distributed under the CC BY-SA license; TyDi QA – under the Apache-2.0 license.

In addition to monolingual test data, we experimented with two parallel/cross-lingual datasets: MLQA and XQuAD and explored two directions: (1) question is in a target language, but paragraph is in English; (2) a question is in English, but a paragraph is in a target language.

QA performance is evaluated using token-level F1-score.

## 5 Experimental Results and Analysis

### 5.1 Setup

All experiments were conducted on a single Tesla V100 16GB. For cross-lingual model adjustment we use the Adam optimizer and hyper-parameters provided by Cao et al. (2020). To obtain reliable results we run five iterations (using different seeds) of model adjustment (for each configuration) followed by fine-tuning on downstream tasks. For each run we sample a required number of sentences from a set of 250K parallel (WikiMatrix) sentences word-aligned with *fast_align*.[6] One run of model adjustment on 30K parallel sentences takes about 15 minutes.

The code to fine-tune mBERT on XNLI, SQuAD, and Wikiann is based on `HuggingFace` sample scripts,[7] which were modified to support continual learning. These scripts use a basic architecture consisting of a BERT model and a task-specific linear layer. We also reuse parameters provided by `HuggingFace`, except for the weight $\alpha = 0.01$ in the multi-task loss (Eq. 2), which was tuned on a validation set. Also note that batch sizes are 32 (for the main target loss) and 16 (for the auxiliary cross-lingual adjustment loss in the case of continual learning). Fine-tuning on XNLI, SQuAD and Wikiann takes about 100, 60, and 3 minutes, respectively. With continual learning it takes 240, 90, and 15 minutes, respectively. Including all preliminary and exploratory experiments the total computational budget was approximately 500 hours.

[6] We ran the main body of experiments with 30K parallel sentences. In addition, we conducted experiments with 5K/10K/30k/100K/250K Ru-En sentence pairs, see § A.1.

[7] https://github.com/huggingface/transformers/tree/master/examples/pytorch

| mBERT | es | ru | vi | hi |
|---|---|---|---|---|
| Original XNLI | | | | |
| Original | 74.20 | 67.95 | 69.58 | 59.03 |
| Adjusted | 74.82* | 69.45* | 70.88* | 61.54* |
| Adjust.+continual | **75.89**** | **71.26**** | **72.79**** | **63.90**** |
| Mixed-language NLI | | | | |
| Original | 70.93 | 64.24 | 62.72 | 53.53 |
| Adjusted | 72.06* | 66.56* | 66.50* | 57.31* |
| Adjust.+continual | **73.50**** | **69.09**** | **69.14**** | **61.09**** |

Statistically significant differences from an original and adjusted mBERT are marked with * and **, respectively (p-value threshold 0.05).

Table 2: Performance on original and mixed-language NLI datasets (accuracy).

| mBERT | es | ru | vi | hi |
|---|---|---|---|---|
| Original | **73.40** | 63.43 | 71.02 | 65.24 |
| Adjusted | 73.28 | 65.49* | **71.99*** | **68.22*** |
| Adjust.+continual | 72.71** | **66.27**** | 71.35** | 66.07** |

Statistically significant differences from an original and adjusted mBERT are marked with * and **, respectively (p-value threshold 0.05).

Table 3: Performance on NER tasks (token-level F1).

All reported results are averages over five runs with different seeds. We further assess significance of differences between results for the original and adjusted mBERT using paired statistical tests. For QA and XNLI we first average metric values for each example over different runs and then carry out a paired t-test using averaged values. For NER we concatenate example-specific predictions for all seeds and run 1,000 iterations of a permutation test for concatenated sequences (Pitman, 1937; Efron and Tibshirani, 1993).

### 5.2 Main Results

Results for NLI, NER, and QA tasks are summarized in Tables 2, 3, and 4, respectively. We can observe consistent and statistically significant improvements (up to 2.5 accuracy point) of aligned models over zero-shot transfer on XNLI for all languages. This is in line with Cao et al. (2020) even though we used a set of more diverse languages, presumably noisier parallel data, and a slightly different learning scheme. Employing continual learning leads to additional substantial gains (up to 2.4 accuracy points). We also evaluated models on the (bilingual) mixed-language XNLI test data (see § 4.2). According to the bottom part of Table 2, compared to the original XNLI, we observe bigger gains for all four languages, especially when we employ continual learning. For Hindi, we obtain a 7.5 point gain by using both the adjustment and continual learning.

| mBERT | Spanish | | Russian | | Vietnamese | | Hindi | |
|---|---|---|---|---|---|---|---|---|
| | MLQA | XQuAD | TyDi QA | XQuAD | MLQA | XQuAD | MLQA | XQuAD |
| Original | **64.96** | **75.59** | 67.05 | **70.72** | **59.95** | **69.18** | **48.73** | 57.56 |
| Adjusted | 63.11* | 73.99* | 67.03 | 70.58 | 58.46* | 68.63 | 48.47 | 57.81 |
| Adjust+continual | 62.76** | 73.44 | **67.63** | 70.51 | 57.71** | 68.64 | 48.02** | **57.83** |
| Question in target language, paragraph in English | | | | | | | | |
| Original | **67.34** | **75.74** | – | 71.54 | 56.08 | 65.00 | 42.48 | 47.83 |
| Adjusted | 66.93* | 75.65 | – | **71.68** | **56.74*** | **66.75*** | **44.91*** | **50.45*** |
| Adjust+continual | 66.31** | 74.88** | – | 70.99 | 54.51** | 64.63** | 43.88** | 50.13 |
| Question in English, paragraph in target language | | | | | | | | |
| Original | **67.36** | **76.71** | – | 67.31 | 64.43 | 68.12 | 55.32 | 58.62 |
| Adjusted | 66.96* | 76.42 | – | **68.25*** | **65.01*** | **68.99** | **55.63** | **58.93** |
| Adjust+continual | 66.68** | 76.21 | – | 68.06 | 64.36** | 68.54 | 54.74** | 58.22 |

Statistically significant differences from an original and adjusted mBERT are marked with * and **, respectively (p-value threshold 0.05).

Table 4: Effectiveness of QA systems (F1-score).

NER results are somewhat mixed: We observe statistically significant gains (up to 3 points for Hindi) on all languages except Spanish. In that, continual learning is beneficial only for Russian.

When we fine-tune a cross-lingually adjusted mBERT on QA tasks, there are no statistically significant gains. In that, there is a statistically significant decrease for all Spanish datasets and Vietnamese MLQA. Use of continual learning leads to further degradation in nearly all cases. Note that models are noticeably more accurate on XQuAD compared to MLQA, which can be due to XQuAD being a translation of SQuAD, which is, in turn, is used to train our QA models.

Muttenthaler et al. (2020) and van Aken et al. (2019) showed that QA models essentially clustered answer token vectors and separated them from the rest of the paragraph token vectors using a vector representation of the question. Thus, to solve the QA task, the model learns to rely on *mutual similarities* among question and answer tokens (on English QA data) rather than on their actual vector representations. As a consequence, there is no need to make representations in the target language to be similar to English-language representations. which, in turn, may *partially* explain why the cross-lingual adjustment is unsuccessful for QA.

We further hypothesizes that such an alignment is more crucial for cross-lingual tasks, which we can be partially corroborated using experiments with two parallel datasets: MLQA and XQuAD. We explored two directions: (1) question is in a target language, but paragraph is in English; (2) a question is in English, but a paragraph is in a target language. According to results in the lower part of Table 4, there are, indeed, several cases when the adjustment is beneficial. Note that cross-lingual adjustment is also more useful for the mixed, i.e., cross-lingual XNLI data, than for original one (Table 2). We observe improvements for Vietnamese and Hindi, which is in line with cross-lingual QA results by Faisal and Anastasopoulos (2021), but there are no gains for Spanish and Russian.

### 5.3 Diagnostic Experiments

The lackluster performance of the cross-lingual adjustment of Cao et al. (2020) on QA data motivated us to carry out additional "diagnostic" experiments: We hope to discover potential issues with our setup or derive additional insights about the problem and data. None of these tests, however, uncovered any anomalies or issues.

First we conjectured that the adjustment somehow harms mono-lingual capabilities of mBERT. Comparing original and adjusted mBERT fine-tuned on original and/or translated SQuAD and tested on SQuAD and MLQA data did not support this conjecture, see § A.2.

Second, we hypothesized that our parallel corpora lacked in either quality or quantity, which we tested on Russian data. The quality was checked by aligning Yandex ru-en corpus,[8] which did not lead to better results compared to WikiMatrix. In § A.1, we showed that as the amount of parallel data increases, this clearly improved both NLI and NER. In that, QA performance peaked at roughly 5K parallel sentences and further decreased as the number of parallel sentences increases.

Third, we ran experiments with an extended version of m-BERT (see § A.3), to ensure that relative (original vs adjusted mBERT) performance on Hindi is not negatively affected by the low qual-

[8] https://translate.yandex.ru/corpus

ity of mBERT token inventory for Hindi. These experiments show that the adjustment can, indeed, improve Hindi QA models.

K et al. (2020) showed that the quality of cross-lingual transfer was higher in the case of languages with similar word order. Hsu et al. (2019) and Zhao et al. (2021) experimented with word rearrangements for cross-lingual QA and NLI, respectively, and obtained some improvements. We trained a QA model using an English-Hindi adjusted mBERT on the SQuAD-SOV dataset released by Hsu et al. (2019), where sentences were re-arranged to Subject-Object-Verb order. This combination led to a degraded quality.[9]

### 5.4 Analysis of the Adjusted mBERT

We calculate $L_2$ distances between contextualized embeddings in English and other languages.[10] The embeddings are taken from the last layer output (i.e., no prediction heads are used). To this end we sampled semantically related words from parallel sentences (matched via *fast_align*) and unrelated words from unpaired sentences (nearly always unrelated). For each pair of languages and each NLP task, the sampling processed is carried out for: (1) the original mBERT, (2) an adjusted mBERT, (3) the original mBERT fine-tuned for the target NLP task, (4) the adjusted mBERT fine-tuned for the target NLP task, (5) the adjusted mBERT fine-tuned for the target task using *continual* learning (full set of histograms can be found in Appendix B).

Timkey and van Schijndel (2021), among others, report that in a *monolingual setting* there are a few (single digit) "rogue" dimensions that dominate computation of the cosine similarity. Yet, these dimensions do not explain model behavior, which makes such distance analysis pointless. This is less problematic on our data with $L_2$ distance: For example for all XNLI models, 10 and 100 most "influential" dimensions account only for about 5% and 25% of the overall distance, respectively.

From Fig. 1 we can see that the cross-lingual adjustment makes embeddings of semantically similar words from different languages closer to each other while keeping unrelated words apart, which is in line with Zhao et al. (2021). However, prior work did not inspect histograms obtained after fine-tuning. Yet, quite surprisingly, fine-tuning of both the original and adjusted mBERT on the English NLI data (Fig. 1c and 1d) makes distributions of related and unrelated words almost fully overlap, i.e. all embeddings become close to each other. Compared to the original mBERT, fine-tuning of the *adjusted* mBERT (Fig. 1d) does result in a better separation of related and unrelated words, but the effect is *quite modest*. We believe this is an example of "catastrophic forgetting" (McCloskey and Cohen, 1989), where fine-tuning the model on a target task causes the model to forget some of the knowledge obtained during cross-lingual adjustment.

Continual learning (Fig. 1e) permits fine-tuning for the target task while maintaining a separation between related and related words, which also consistently improves performance for the NLI task. It is crucial to note that we see no direct relationship between the degree of separation and the success of cross-lingual transfer among all tasks and training regimes. In the case of NER, the biggest separation is achieved for Spanish (see Fig. 2 in Appendix B), but fine-tuning of the adjusted mBERT results in a lower accuracy. More generally, fine-tuning with continual learning *always* leads to better separation of related and unrelated words, but this is beneficial only for the NLI task.

## 6 Conclusion

We evaluate effectiveness of an existing approach to cross-lingual adjustment of mBERT (Cao et al., 2020) using four typologically different languages (Spanish, Russian, Vietnamese, and Hindi) and three NLP tasks (QA, NLI, and NER). The original mBERT is being compared to mBERT "adjusted" with a help of a small parallel corpus. The cross-lingual adjustment of mBERT improves NLI in four languages and NER in three languages. However, in the case of QA performance never improves and sometimes degrades. For Hindi data, this happens due to a lower quality of mBERT on Hindi data. Inspired by the analysis of histograms of distances, we obtain *additional* improvement on NLI using continual learning. Our study contributes to a better understanding of cross-lingual transfer capabilities of large multilingual language models. It also identifies limitations of their cross-lingual adjustment in various NLP tasks.

[9] Manual inspection of the data revealed that all SQuAD data is lowercased, which may negatively impact QA training. Moreover, the quality of rearrangements is rather low, most obvious problem is incorrect processing of passive voice constructions.

[10] Although most prior work uses the cosine similarity instead of $L_2$ (Rudman et al., 2022), it does not distinguish between vectors with the same direction, but different lengths.

## References


Hanan Aldarmaki and Mona Diab. 2019. Context-aware cross-lingual mapping. 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 3906–3911.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *ACL*, pages 4623–4637.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *8th International Conference on Learning Representations, ICLR 2020*.

Rich Caruana. 1996. Algorithms and applications for multitask learning. In *ICML*, pages 87–95. Morgan Kaufmann.

Ethan A Chi, John Hewitt, and Christopher D Manning. 2020. Finding universal grammatical relations in multilingual bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577.

Jonathan H Clark et al. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *TACL*, 8:454–470.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel 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*, pages 2475–2485.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. 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.

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

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, et al. 2021. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. *arXiv preprint arXiv:2104.08726*.

Bradley Efron and Robert Tibshirani. 1993. *An Introduction to the Bootstrap*. Springer.

Fahim Faisal and Antonios Anastasopoulos. 2021. Investigating post-pretraining representation alignment for cross-lingual question answering. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 133–148.

Tsung-Yuan Hsu, Chi-Liang Liu, and Hung-yi Lee. 2019. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5933–5940.

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.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: an empirical study. In *ICLR*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, pages 79–86.

Saurabh Kulshreshtha, Jose Luis Redondo Garcia, and Ching-Yun Chang. 2020. Cross-lingual alignment methods for multilingual BERT: A comparative study. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 933–942.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *ACL*, pages 7315–7330.

Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert? *arXiv preprint arXiv:1911.03310*.

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

Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2021. Evaluating multilingual text encoders for unsupervised cross-lingual retrieval. *arXiv preprint arXiv:2101.08370*.

Zihan Liu, Genta Indra Winata, Andrea Madotto, and Pascale Fung. 2021. Preserving cross-linguality of pre-trained models via continual learning. In *RepL4NLP@ACL-IJCNLP*, pages 64–71. Association for Computational Linguistics.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462.

Lukas Muttenthaler, Isabelle Augenstein, and Johannes Bjerva. 2020. Unsupervised evaluation for question answering with transformers. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 83–90.

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.

German Ignacio Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71.

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.

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.

Edwin JG Pitman. 1937. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Roger Ratcliff. 1990. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285.

Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2021. Synthetic data augmentation for zero-shot cross-lingual question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7016–7030.

Anthony Robins. 1995. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2021. QA dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *arXiv preprint arXiv:2107.12708*.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. 2022. Isoscore: Measuring the uniformity of embedding space utilization. In *ACL*

*(Findings)*, pages 3325–3339. Association for Computational Linguistics.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. 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 1599–1613.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361.

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

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *EMNLP (1)*, pages 4527–4546. Association for Computational Linguistics.

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*.

Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A Gers. 2019. How does BERT answer questions? A layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1823–1832.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019a. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5721–5727, Hong Kong, China.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime Carbonell. 2019b. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. *arXiv preprint arXiv:1910.04708*.

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

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)*, pages 833–844.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing language-agnostic multilingual representations. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 229–240.

# A Additional Experiments

## A.1 Hyper-Parameter Tuning

An objective of this section analysis is to assess the impact of most important hyper-parameters, which might have substantially affected outcomes. We focus on the size of the parallel corpus, on the approach to aggregating subword embedding, and a choice of aligned tokens (is using `[CLS]` and/or `[SEP]` beneficial?). In all cases we measure a performance gain/loss compared to the original, i.e., unadjusted mBERT.

| Size | XNLI | NER | TyDi QA | XQuAD |
|---|---|---|---|---|
| 5K | +0.62 | +2.06 | +0.24 | +0.75 |
| 10K | +0.76 | +2.01 | +0.62 | +0.42 |
| 30K | +1.50 | +2.07 | –0.02 | –0.14 |
| 100K | +2.24 | +2.19 | –1.09 | –0.11 |
| 250K | +2.67 | +2.53 | –1.98 | –2.17 |

Table 5: Performance gains (compared to the original mBERT) of models aligned on En-Ru data depending on the number of sentence pairs.

**Size and quality of the parallel corpus.** Because zero-shot transfer is typically more challenging for languages with non-Latin script: compare results for Spanish and Vietnamese vs. results for Russia and Hindi in Tables 2 and 3, we initially considered experimenting with either Russian or Hindi. Eventually, we chose Russian, because the Russian Wikipedia is much larger compared to Hindi. As a result it has a better alignment quality as indicated by higher margin scores (Schwenk et al., 2021). In addition of WikiMatrix, we experimented with the Yandex ru-en corpus,[11] but it did not produce better results compared to WikiMatrix.

We adjusted mBERT on several parallel corpora where the number of paired sentences ranged from 5K to 250K. We then fine-tuned the *adjusted* model for several target tasks. As in all other experiments, we train the models with five seeds and report averaged results. Table 5 shows that XNLI and NER accuracy improves nearly monotonically as the size of the parallel corpus increases.

QA models benefit from adjustment using only a small amount of parallel data (and even slightly outperform the original mBERT baseline when adjusted using 5K sentence pairs). QA performance peaks at roughly 5K parallel sentences and further decreases as the number of parallel sentences increases. This seems to be some form of overfitting, but the reasons are unclear: We tried to carry out a cross-lingual adjustment with the learning rates inversely proportional to the parallel corpus size, but the improvements were small and inconsistent.

[11] https://translate.yandex.ru/corpus

| Mode | XNLI | NER | MLQA | XQuAD |
|---|---|---|---|---|
| start | +2.36 | +3.08 | –0.16 | +0.01 |
| end | +2.39 | +2.59 | –1.12 | –0.44 |
| avg | +2.51 | +2.98 | –0.25 | +0.24 |
| subword alignment | +2.37 | +0.99 | –5.01 | –4.42 |

Table 6: Impact of subword aggregation approach (Hindi): Performance gains compared to the original mBERT.

| Adjustment by | XNLI | NER | MLQA | XQuAD |
|---|---|---|---|---|
| [CLS] | +1.15 | +1.61 | –0.12 | +0.5 |
| [CLS] [SEP] | +1.25 | +1.83 | –0.50 | +0.2 |
| words | +2.48 | +3.03 | –0.26 | –0.03 |
| all | +2.51 | +2.98 | –0.25 | +0.24 |

Table 7: Impact of special and word tokens (Hindi): Performance gains compared to the original mBERT.

**Subword embedding aggregation.** In our main experiments, we align words by using their averaged (*avg*) subword embeddings, which performed best in preliminary QA experiments. However, as Table 6 shows this is not an optimal approach across all tasks and languages. For example, in the case of Hindi, we get better results using the first token (*start*) on NER task (though differences are small).

Interestingly, when we apply *fast_align* to original WordPiece tokens (*subword alignment*), we obtain much worse results on all tasks except NLI. We hypothesize that a lower quality of the *subword alignment* approach is likely due to a small mBERT vocabulary allocated for Hindi. This leads to excessive word splitting and, consequently, to a worse alignment. We confirm this conjecture in § A.3.

**A choice of aligned tokens.** In our main experiments, the mBERT adjustment procedure uses both regular word and special-word tokens `[CLS]` and `[SEP]`. In Table 7 we show ablation experiments where we exclude some of the tokens from the alignment procedure. We conjectured that in the NLI task the model relies more on the sentence-level representation through a `[CLS]` token. However, a more then one-point gain is achieved by aligning only words, which is slightly improves when the alignment additionally uses `[CLS]`. The same is true for the NER task. In the case of QA, aligning only the `[CLS]` token is suboptimal, but

| mBERT | SQuAD | | | | MLQA | | |
|---|---|---|---|---|---|---|---|
| | es | ru | vi | hi | es | vi | hi |
| Original | **89.26** | | | | **66.63** | **66.15** | 60.64 |
| Adjusted | 89.0 | 88.99 | 88.95 | 89.0 | 66.52 | 66.06 | **60.77** |

Table 8: Comparison of original mBERT and aligned mBERT on mono-lingual QA data (F1-score). First row after the caption shows languages used in cross-lingual alignment (and the language of the dataset for MLQA). Models tested on MLQA are trained on the translated SQuAD.

| EmBERT | XNLI | NER | MLQA | XQuAD |
|---|---|---|---|---|
| Original | 63.76 | 65.12 | 53.47 | 64.11 |
| Adjusted by 30K | 65.70 | **66.79** | **54.94** | 64.11 |
| Adjusted by 100K | **66.69** | 66.29 | 54.66 | **65.10** |

Table 9: Extended mBERT for Hindi.

combining regular words with special tokens is beneficial.

### A.2 Assessing Mono-lingual Capabilities of Adjusted mBERT

The main objective of this section is to assess if the mono-lingual capabilities of mBERT were negatively affected by the cross-lingual adjustment. To this end, we fine-tune adjusted and the original mBERT on monolingual QA data, which includes the original SQuAD dataset as well as its *machine translations* into three languages: Spanish, Vietnamese, and Hindi released along the MLQA dataset.[12]

The results are shown in Table 8, where the first row after the caption shows languages used in cross-lingual alignment. In the case of translated SQuAD a language used in in the adjustment coincides with the evaluation language. For SQuAD evaluation is done in English. According to Table 8, there are only marginal (at most 0.3%) differences between the F1-scores of original and adjusted mBERT.

### A.3 Evaluating Extended mBERT for Hindi

We were concerned that results on Hindi were affected by the poor quality of the token inventory, which—compared to English—leads to a substantial word segmentation. Thus, we carried out additional experiment using an "extended" mBERT version, which has a much better token inventory for Hindi (Wang et al., 2020).[13] Comparing Tables 9 and Tables 2, 3, 4, we can see that using the extended mBERT allows us to achieve better results. Moreover, applying the cross-lingual adjustment produces at least one point gain for all tasks including both MLQA and XQuAD. We, thus conclude that poor performance of the adjustment procedure on Hindi data can be attributed to a lower quality of mBERT on Hindi data.

[12] https://github.com/facebookresearch/MLQA

[13] https://github.com/ZihanWangKi/extend_bert

## B Histograms of $L_2$ Distances between Contextualized Embeddings

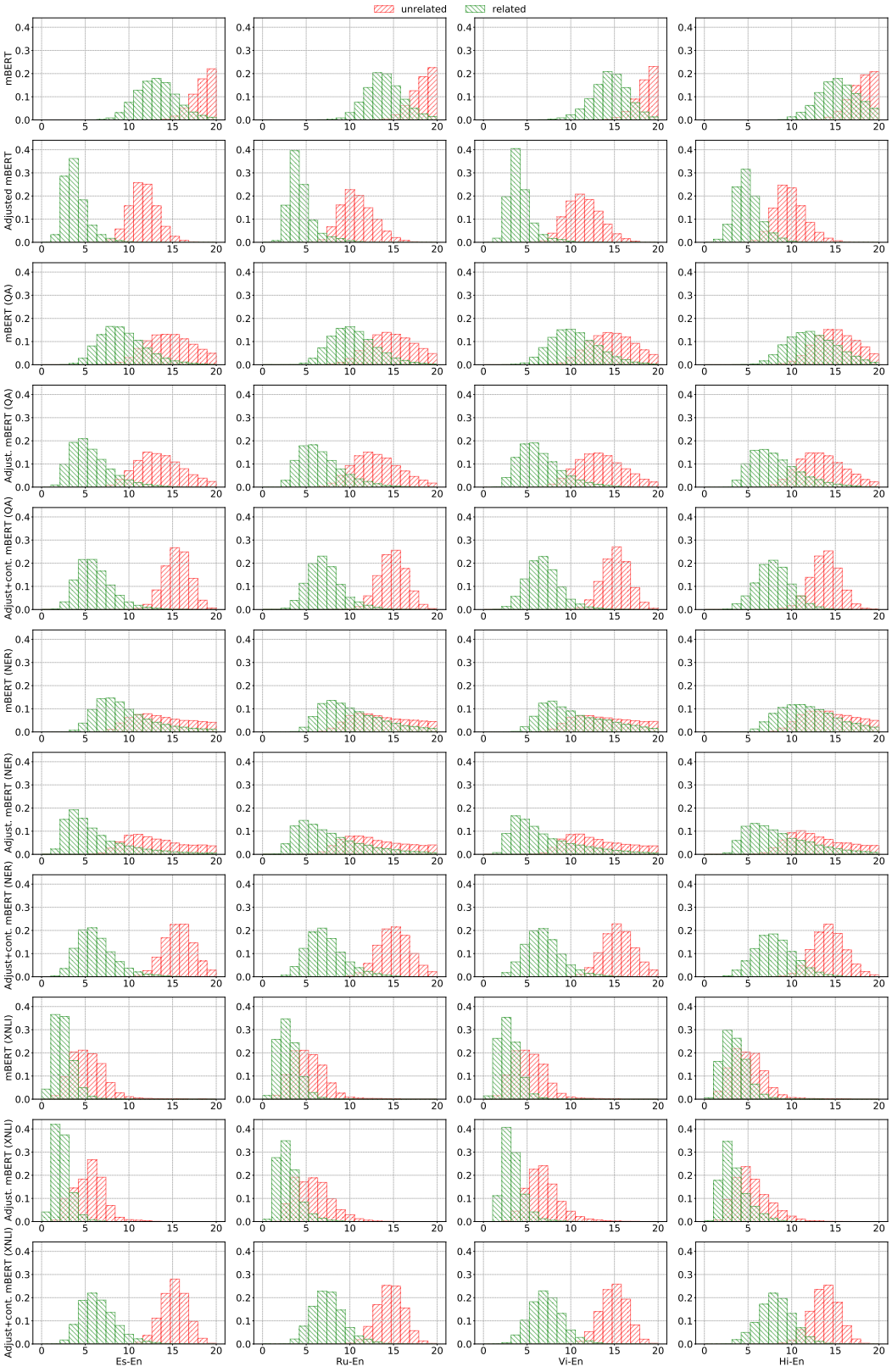


Figure 2: Histograms of $L_2$ distances between pairs of contextualized representations (produced by mBERT) for randomly sampled related (i.e., aligned) and unrelated word pairs from WikiMatrix. Columns correspond to language pairs. Rows depict histograms of the original mBERT model, its cross-lingual adjustments, as well as their variants fine-tuned on QA, NER, and NLI tasks using a regular as well as a continual mode.