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

Anonymous ACL submission

Abstract

Large pre-trained multilingual models such as mBERT and XLM-R enabled effective cross-lingual zero-shot transfer in many NLP tasks. A cross-lingual adjustment of these models using a *small* parallel corpus can further improve results. This is a more data efficient method compared to training a machine-translation system or a multi-lingual model from scratch using only parallel data. In this study, we experiment with zero-shot transfer of English models to four typologically different languages (Spanish, Russian, Vietnamese, and Hindi) and three NLP tasks (QA, NLI, and NER). We carry out a cross-lingual adjustment of an off-the-shelf mBERT model. We show that this adjustment makes embeddings of semantically similar words from different languages closer to each other, while keeping unrelated words apart. In contrast, fine-tuning of mBERT on English data (for a specific task such as NER) draws embeddings of *both* related and unrelated words closer to each other. The cross-lingual adjustment of mBERT improves NLI in four languages and NER in two languages. However, in the case of QA performance never improves and sometimes degrades. In that, the increase in the amount of parallel data is most beneficial for NLI, whereas QA performance peaks at roughly 5K parallel sentences and further decreases as the number of parallel sentences increases.

1 Introduction

Natural disasters, military operations, or disease outbreaks require a quick launch of information systems relying on human language technologies. Such systems need to provide instant situational awareness based on sentiment analysis, named entity recognition (NER), information retrieval, and question answering (QA) (Roussinov et al., 2008; Voorhees et al., 2020; Chan and Tsai, 2019; Strassel and Tracey, 2016). The quality of these techniques heavily depends on the existence of annotated data, which is particularly challenging in low-resource languages. Large langauge models pre-trained on a large multilingual corpus such as mBERT or XLM-R enable a zero-shot *cross-lingual* transfer by learning to produce contextualized word representations, which are (to some degree) language-independent (Libovickỳ et al., 2019; Pires et al., 2019). These representations can be further aligned using a modest amount of parallel data, which was shown to improve zero-shot transfer for syntax 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 most common approach 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, following (Cao et al., 2020) 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. Then, we fine-tune cross-lingually adjusted mBERT models for three NLP tasks (NLI, NER, and QA) using English data. 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

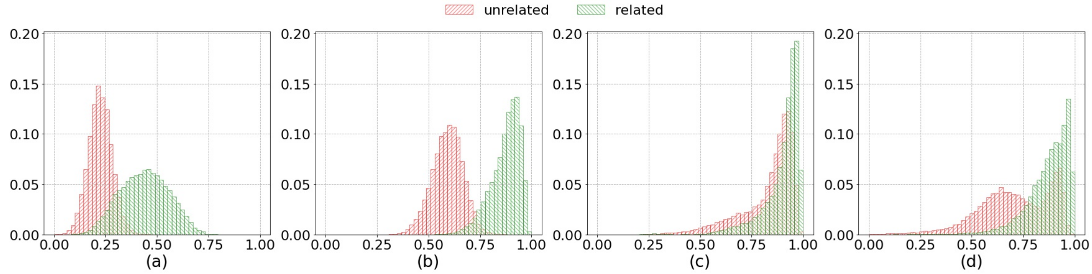

Figure 1: Histograms of cosine similarities between contextualized word representations produced by mBERT for 20,000 randomly sampled (unrelated) vs. aligned (related) word pairs from WikiMatrix (Hi-En): (a) original, (b) after cross-lingual adjustment, (c) after fine-tuning on English NLI data, (d) after cross-lingual adjustment and subsequent fine-tuning on English NLI data.

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 subsequently fine-tuned on English data and zero-shot transferred to a target language perform on various NLP tasks and target languages?

R2 How do the size of the parallel corpus used for adjustment and different approaches to word alignment affect outcomes?

R3 How do 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?

Our experiments demonstrate the following:

- The cross-lingual adjustment of mBERT improves NLI in four languages (by one point) and NER in two languages (by 1.5-2.5 points). Yet, there is no statistically significant improvement for QA and a statistically significant deterioration on three out of eight QA datasets.

- Although a choice of a word-alignment approach (e.g., averaging over word sub-tokens) slightly affects outcomes, there are no apparent patterns. However, as the amount of parallel data increases, this clearly 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 similarity of contextualized-embeddings of words across languages (Fig. 1), we can see that the cross-lingual adjustment of mBERT increases the cosine similarity between related words while keeping unrelated words apart. In contrast, fine-tuning of mBERT for a specific task draws embeddings of *both* related and unrelated words much closer to each other (Fig. 1c). However, when we fine-tune a cross-lingual adjusted mBERT for a specific task (e.g., NLI), cosines similarities between related and unrelated words are better separated (Fig. 1d), which may permit the adjusted mBERT to have better zero-shot transfer performance.

In summary, our study contributes to a better understanding of large multilingual language models and their cross-lingual transfer capabilities by identifying limitations of this approach in various NLP tasks. To enable reproducibility, we make our software available (currently attached to the submission).

2 Related Work

2.1 Cross-Lingual Zero-Shot Transfer with Multilingual Models

The success of mBERT in cross-language zero-shot regime on many tasks 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. As a result, transfer is 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 includes NLI, NER, and QA datsets used in the current study. Authors state that performance on question answering on XTREME has improved only slightly since its inception in contrast to a more impressive progress in e.g. classification and retrieval tasks. 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. demonstrated that vector spaces can encode semantic relationships between words and that there are similarities in the geometry of these vectors spaces across languages (Mikolov et al., 2013). 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 is 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 method 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 making the representations of semantically related words in different languages to be closer to each other. This work 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 mBERT simultaneously adjusted on five European languages consistently outperformed other alignment approaches on XNLI data (Conneau et al., 2018). In the current study, we implement the approach of Cao et al. (2020) with some modifications.

Kulshreshtha et al. (2020) compared different alignment methods (rotation vs. adjustment). They evaluated the modified embeddings on NER and slot filling tasks. According to their results, rotation-based alignment performs better on 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 alignment 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.

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 masked language model objective on a mixture of Wikipedias of 104 languages with a shared WordPiece vocabulary: To balance the distribution of languages, high-resource languages were under-sampled and low-resource languages were over-sampled.[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 (see Section 2).

We align cross-lingual embeddings by directly modifying/adjusting the language model itself, following the approach proposed recently 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, which splits sufficiently long words into subtokens. We first word-align parallel data with *fast_align* (Dyer et al., 2013) and then employ three common approaches to combine subtoken vectors into a single vector representing a word: 1) using the vector of the first sub-token[3]; 2) using the vector of the last subtoken (Cao et al., 2020); 3) averaging of all word subtokens. We also explored another variant: applying *fast_align* directly to BERT tokenization (i.e., subtokens). We use the *averaging* approach for our main experiments. Additionally, we assess how the choice of the alignment approach affects performance on Hindi data.

Lang	Family	Script	Word order	Dist. from English	Number of Wiki pages
en	IE/Germanic	Latin	SVO	0.00	6.3M
es	IE/Romance	Latin	SVO	0.12	1.7M
ru	IE/Slavic	Cyrillic	SVO	0.14	1.7M
vi	Austroasiatic	Latin	SVO	–	1.3M
hi	IE/Indo-Aryan	Devanagari	SOV	0.40	150K

IE : Indo-European; Prevalent word order: SVO – subject-verb-object, SOV – subject-object-verb; distance from English in terms of word order is measured according to Ahmad et al. (2019): Data for Vietnamese is missing.

Table 1: Language information.

Based on alignments in parallel data, we obtain a collection of word or subtoken (depending on the processing variant, see above) 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 fine-tune the mBERT model on aligned pairs' vectors using the following loss function:

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

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 $\mathbf{f}^0$.

After have cross-lingually adjusted the mBERT model, we fine-tune it for a specific task.

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 (although Hindi Wikipedia is an order of magnitude smaller compared to other Wikipedias, which may have led to somewhat inferior contextualized embeddings).

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. Following (Wang et al., 2019b; Kulshreshtha et al., 2020), we take 30K sentence pairs for each language pair as a 'basic' size.[4]

4.2 Natural Language Inference

Natural language inference (NLI) is task of determining the relation between two ordered sentences and classifying it into: entailment, contradiction, or "no relation". English MultiNLI collec-

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

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

[3] Wang et al. (2019b) used this variant in their experiments and report that other options don't induce much difference.

[4] Cao et al. (2020) use the same magnitude of data – 50K sentence pairs for each out of five languages.

tion (Williams et al., 2018) consists of 433K sentence pairs originating from multiple genres. The XNLI dataset (Conneau et al., 2018) complements the MultiNLI training set with 2,500 development and 5,000 test examples in each of 15 languages (including all four target languages of the current study). Performance on XNLI is evaluated using classification *accuracy*. XNLI is distributed under the CC BY-NC license.

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. To create SQuAD, crowd workers were shown a Wikipedia paragraph; the task was to formulate several questions to the paragraph content and select a text span as an answer. SQuAD is available under the CC BY-SA license. SQuAD has become a *de facto* standard and inspired creation of analogous resources in other languages (Rogers et al., 2021).

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 English). 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). Compared to datasets associated with SQuAD, SQuAD, XQuAD, and MLQA are distributed under the CC BY-SA license; TyDi QA – under the Apache-2.0 license.

Standard evaluation metrics for SQuAD-like datasets are EM (exact answer-span match) and token-level F1-score. We report F1-scores because they are considered to be more robust.

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 hyperparameters 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 down-stream tasks. For each run we sample a required number of sentences from a set of 250K parallel (WikiMatrix) sentences word-aligned with *fast_align*. One run of model adjustment on 30K parallel sentences takes about 15 minutes.

For fine-tuning on XNLI, SQuAD, and Wikiann we use parameters and scripts provided by `HuggingFace`.[5] These scripts use a basic architecture consisting of a BERT model and a task-specific linear layer. We freeze the embedding vectors during fine-tuning on down-stream tasks because during the training on English data the model ignores vectors in other languages. It can also prevent forgetting of embedding adjustment. Fine-tuning on XNLI, SQuAD and Wikiann takes about 100 minutes, 60 minutes, and 3 minutes respectively. Including all preliminary and exploratory experiments the total computational budget was approximately 450 hours.

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

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

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 1000 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 of aligned models over zero-shot transfer on XNLI for all languages. All gains are around one accuracy point, which is in accordance with results by Cao et al. (2020) even though we used a set of more diverse languages, presumably noisier parallel data, and slightly altered learning scheme. Results confirm previous findings about cross-lingual zero-short transfer: results are higher for cognate target languages (cf. Spanish) and become worse as you move farther away from the source language (Hindi demonstrates worst results, but the deficiency of Hindi data for mBERT learning may also come into play). We also constructed bilingual variant of XNLI test data: in each pair, we randomly swapped one of the sentences with its English counterpart. Classification results for this "mixed" dataset in the bottom of Table 2 demonstrate a larger gain of the adjusted mBERT (compared to the original mBERT) for all four languages.

NER results are mixed: we observe significant gains for Russian (+2.3 F1 points) and Hindi (+2.6 F1 points) when using a cross-lingually adjusted model, while the results for Spanish and Vietnamese are worse compared to the original mBERT. At the same time, the baseline scores for Russian and Hindi are lower compared to English and Vietnamese. We hypothesize that the reason may be the annotation quality of the Wikiann corpus, which varies across languages.

When we fine-tune a cross-lingually adjusted mBERT on QA tasks, we observe a statistically significant performance degradation for both Spanish datasets as well as for Vietnamese MLQA. Again, we observe a steady decline of zero-shot transfer outcomes on both parallel datasets (MLQA and XQuAD) from languages closer to English (Spanish) to more distant ones (Hindi). It is also interesting to note the gap between the MLQA and XQuAD scores. XQuAD is a translation of a SQuAD and, therefore, it "inherits" a higher lexical similarity of the question and answer contexts. This makes it an easier task for a model fine-tuned on the original SQuAD.

mBERT	es	ru	vi	hi
Original	74.59	68.26	70.29	59.64
Adjusted	75.52*	69.39*	71.21*	60.77*
Mixed-language NLI				
Original	71.22	65.02	63.12	54.16
Adjusted	73.10*	67.47*	67.29*	57.51*

Statistically significant differences between original and adjusted mBERT are marked with * (p-value threshold 0.05).

Table 2: NLI results (accuracy).

Model	es	ru	vi	hi
Original	73.33	64.53	71.71	65.54
Adjusted	72.02*	66.80*	71.08*	68.11*

Statistically significant differences between original and adjusted mBERT are marked with * (p-value threshold 0.05).

Table 3: NER results (token-level F1).

We also conducted experiments on cross-lingual question answering using two parallel datasets: MLQA and XQuAD: Results are shown in the lower part of Table 4. We explored two directions: 1) question is in a target language, but paragraph is in English and 2) a question is in English, but a paragraph is in a target language. Again, results are not consistent across languages and cross-lingual "directions". In most cases the differences between results obtained using the original and adjusted mBERT are not statistically significant.

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.[6]

5.3 Analysis of the Adjusted mBERT

Liu et al. (2021) observed that after fine-tuning on a specific task (POS-tagging and NER) large multilingual models became worse at the tasks they

[6] 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.

mBERT	Spanish		Russian		Vietnamese		Hindi	
	MLQA	XQuAD	TyDi QA	XQuAD	MLQA	XQuAD	MLQA	XQuAD
Original	65.07	75.62	66.74	71.14	60.18	69.42	49.05	57.63
Adjusted	63.96*	74.59*	66.68	70.57	59.34*	69.97	48.81	57.64
Question in target language, paragraph in English								
Original	68.17	76.36	–	72.10	55.79	64.56	43.28	47.74
Adjusted	67.84	76.22	–	72.17	56.72*	66.73*	44.36*	50.28*
Question in English, paragraph in target language								
Original	67.38	76.52	–	67.70	64.27	68.69	55.45	58.44
Adjusted	66.99*	76.71	–	68.05	65.04*	68.86	55.41	58.10

Statistically significant differences between original and adjusted mBERT models are marked with * (p-value threshold 0.05).

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

orig. mBERT	en-es	en-ru	en-vi	en-hi
60.72	60.03	59.45	58.72	60.22

Table 5: SCWS correlation scores of mBERT models: original vs. cross-lingually adjusted on 4 language pairs.

were initially trained for (e.g., predicting a masked word). They also became worse at cross-lingual sentence retrieval. This result motivated us to study how cross-lingual adjustment of mBERT affected the ability of the model to capture semantic similarity in a mono-lingual and cross-lingual settings.

For the mono-lingual, English, evaluation, we used the Stanford Contextual Word Similarity (SCWS) dataset (Huang et al., 2012). It contains contexts for around 2K word pairs along with crowdsourced (ground-truth) similarity ratings for each pair, which allows us to rank them. We compared these ground-truth rankings of word pairs with rankings produced by an original or adjusted mBERT. To this end, a similarity score of two words was computed using the cosine similarity between words' contextualized embeddings. Agreement with ground-truth data was computed using the Spearman's rank correlation.

According to results in Table 5, the cross-lingual adjustment does hurt monolingual performance of mBERT, which is in line with Liu et al. (2021). With an exception of Hindi, the degradation is smaller for Russian and Spanish, which are more closely related to English.

In the cross-lingual evaluation we compared cosine similarities between contextualized embeddings in English and other languages. To this end we sampled from WikiMatrix (Schwenk et al., 2021) using two scenarios: semantically related words from parallel sentences (matched via *fast_align*) and unrelated words sampled 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.

The histograms of cosine similarities for Hindi and NLI task is shown is shown in Figure 1: Histograms for other languages can be found in Appendix A. Figure 1a shows that related words in two languages are typically closer to each other than to randomly selected unrelated words, but the histograms overlap. The cross-lingual adjustment (see Figure 1b) makes embeddings of semantically similar words from different languages closer to each other, while keeping unrelated words apart, which is a desired behavior. In contrast, fine-tuning *original*, i.e., unadjusted, mBERT on the English NLI data (Figure 1c) makes distributions of related and unrelated words almost fully overlap, i.e. all embeddings become close to each other. In that, if we fine-tune the *adjusted* mBERT (Figure 1d), this also reduces the gap between related and unrelated words, but it remains larger compared to fine-tuning of the unadjusted mBERT. Thus, unlike English-only fine-tuning, the cross-lingual adjustment does reduce the similarity gap between related words (from different languages) while keeping unrelated words largely apart. Quite surprisingly, achieving this objective does not seem to be sufficient for improving zero-shot transfer. For example, judging from histograms for NER (see Fig. 2, Appendix A), one of the biggest improvements should be in the case of Spanish, where the cross-lingual adjustment substantially degrades the F1-score.

We can see that the cross-lingual adjustment does not consistently improve QA and NER tasks. It is quite possible that degradation in monolingual performance (as shown using SCWS data, Table 5) is partially responsible for this underwhelming performance, especially for QA, whose quality degrades as the amount of parallel data used for adjustment increases (see Table 6). Furthermore, 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 needs to operate largely at a lexical level and can rely on mutual similarity among question and paragraph words. It learns how to use these similarities by training on the English QA data and does not benefit much from cross-lingual adjustment.

5.4 Impact of the Amount of Parallel Data/Approach to Subtoken Aggregation

An objective of this section analysis is to assess the impact of the number of parallel sentences and the approach to subtoken aggregation. Because zero-shot transfer is typically more challenging for languages with non-Latin script (see Tables 2 and 3), we initially considered experimenting with either Russian or Hindi. Ultimately we chose Russian, because the Russian Wikipedia is much larger compared to Hindi, which entails a higher quality of sentence alignment in WikiMatrix (see Section 4.1).

We adjusted mBERT on 5K/10K/30k/100K sentence pairs and subsequently fine-tuned the model on respective tasks. As in all other experiments, we train the models with five seeds and report averaged results. Table 6 shows that XNLI accuracy improves monotonically as the size of the parallel corpus increases. NER scores reach a plateau after 10K sentence pairs. 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.

Size	XNLI	NER	TyDi QA	XQuAD
None	68.26	64.53	66.74	71.14
5K	68.73	66.02	67.32	71.29
10K	69.38	66.52	67.55	70.66
30K	69.39	66.80	66.68	70.57
100K	70.34	66.80	66.58	69.96

Table 6: Performance of models aligned on En-Ru data depending on the number of sentence pairs.

In our main experiments, we carry out alignment using averaged subtoken embeddings, which was decided based on preliminary experimental results. However, as Table 7 shows, this is not an optimal approach across all tasks and languages. For example, in the case of Hindi, we get better results using start subtokens on NLI and NER tasks (though differences are small for NER).

Interestingly, when we apply *fast_align* to original BERT subtokens (*orig*), we obtain much worse results on all tasks except NLI. A lower quality of the *orig* approach is likely due to a small mBERT vocabulary allocated for Hindi, which results in excessive word splitting and, consequently, leads to a worse alignment. We conjecture that in the NLI task the model relies more on the sentence-level representation through a `[CLS]` token, which are also being aligned as part of the cross-lingual adjustment of mBERT. Good cross-lingual NLI performance with the *orig* approach supports this hypothesis.

Mode	XNLI	NER	MLQA	XQuAD
start	61.59	68.41	48.48	57.16
end	61.24	68.10	47.50	56.46
avg	60.77	68.11	48.81	57.64
orig	61.53	64.54	44.34	53.44

Table 7: Impact of subtokens processing (Hindi).

6 Conclusion

In this study, we experiment with zero-shot transfer of English models to four typologically different languages and three NLP tasks. The cross-lingual adjustment of mBERT improves NLI in four languages and NER in two languages. However, in the case of QA performance never improves and sometimes degrades. Our study contributes to a better understanding of large multilingual language models and their cross-lingual transfer capabilities by identifying limitations of this approach.

References

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on 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 2440–2452.

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.

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*.

Hao-Yung Chan and Meng-Han Tsai. 2019. Question-answering dialogue system for emergency operations. *International Journal of Disaster Risk Reduction*, 41:101313.

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.

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

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.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882.

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 *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 64–71.

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

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.

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.

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*.

Dmitri Roussinov, Weiguo Fan, and José Robles-Flores. 2008. Beyond keywords: Automated question answering on the web. *Communications of the ACM*, 51(9):60–65.

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

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.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280.

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*.

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.

Ellen M. Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2020. TREC-COVID: constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1):1:1–1:12.

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.

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.

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 Appendix

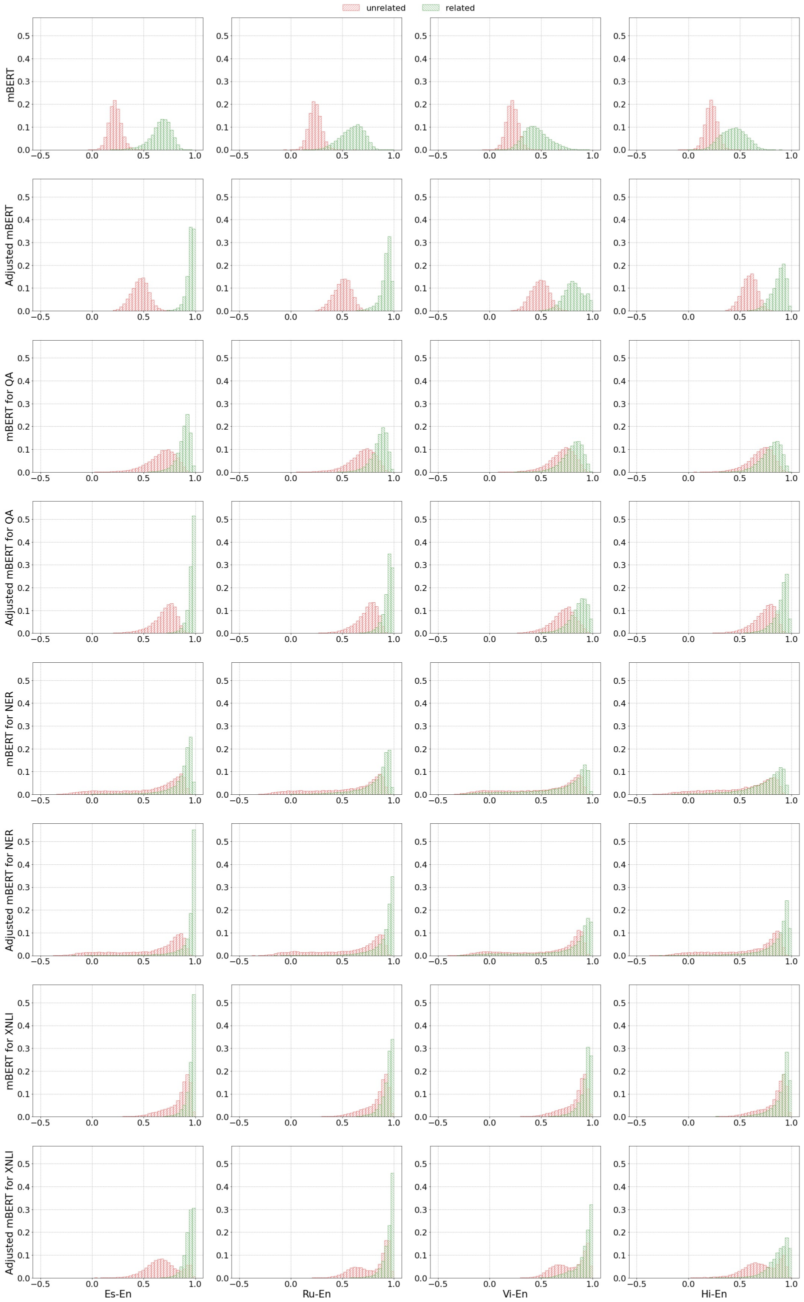

Figure 2: Histograms of cosine similarities between contextualized word representations produced by mBERT for 20,000 randomly sampled (unrelated) vs. aligned (related) word pairs from WikiMatrix. Columns correspond to language pairs. Rows depict histograms of the original mBERT model, its cross-lingual adjustments, as well as their variants fine-tuned on QA, NER, and NLI tasks.