with $U\Sigma V^T = \text{SVD}(YX^T)$, where X and Y are representation of two monolingual BERT models, sampled at different granularities as described below. We apply iterative normalization on X and Y before learning the mapping (Zhang et al., 2019).

5.1.1 Word-level alignment

In this section, we align both the non-contextual word representations from the embedding layers, and the contextual word representations from the hidden states of the Transformer at each layer.

For non-contextualized word embeddings, we define X and Y as the word embedding layers of monolingual BERT, which contain a single embedding per word (type). Note that in this case we only keep words containing only one subword. For contextualized word representations, we first encode 500k sentences in each language. At each layer, and for each word, we collect all contextualized representations of a word in the 500k sentences and average them to get a single embedding. Since BERT operates at the subword level, for one word we consider the average of all its subword embeddings. Eventually, we get one word embedding per layer. We use the MUSE benchmark (Conneau et al., 2017), a bilingual dictionary induction dataset for alignment supervision and evaluate the alignment on word translation retrieval. As a baseline, we use the first 200k embeddings of fastText (Bojanowski et al., 2017) and learn the mapping using the same procedure as §5.1. Note we use a subset of 200k vocabulary of fastText, the same as BERT, to get a comparable number. We retrieve word translation using CSLS (Conneau et al., 2017) with K=10.

In Figure 4, we report the alignment results under these two settings. Figure 4a shows that the subword embeddings matrix of BERT, where each subword is a standalone word, can easily be aligned with an orthogonal mapping and obtain slightly better performance than the same subset of fastText. Figure 4b shows embeddings matrix with the average of all contextual embeddings of each word can also be aligned to obtain a decent quality bilingual dictionary, although underperforming fastText. We notice that using contextual representations from higher layers obtain better results compared to lower layers.

5.1.2 Contextual word-level alignment

In addition to aligning word representations, we also align representations of two monolingual BERT models in contextual setting, and evaluate performance on cross-lingual transfer for NER and parsing. We take the Transformer layers of each monolingual model up to layer i, and learn a mapping W from layer i of the target model to layer i of the source model. To create that mapping, we use the same Procrustes approach but use a dictionary of parallel contextual words, obtained by running the fastAlign (Dyer et al., 2013) model on the 10k XNLI parallel sentences.

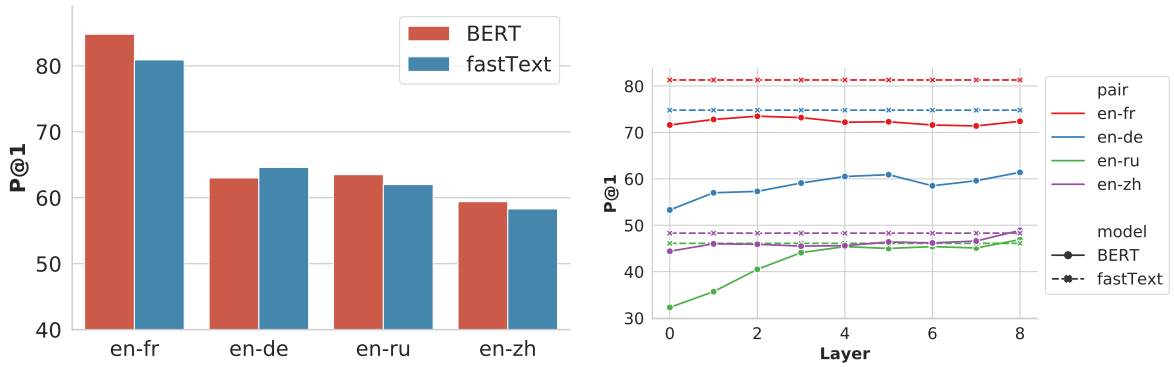
For each downstream task, we learn task-specific layers on top of i-th English layer: four Transformer layers and a task-specific layer. We learn these on the training set, but keep the first i pretrained layers freezed. After training these task-specific parameters, we encode (say) a Chinese sentence with the first i layers of the target Chinese BERT model, project the contextualized representations back to the English space using the W we learned, and then use the task-specific layers for NER and parsing.

In Figure 5, we vary i from the embedding layer (layer 0) to the last layer (layer 8) and present the results of our approach on parsing and NER. We also report results using the first i layers of a bilingual MLM (biMLM). [2] We show that aligning monolingual models (MLM align) obtain relatively good performance even though they perform worse than bilingual MLM, except for parsing on English-French. The results of monolingual alignment generally shows that we can align contextual representations of monolingual BERT models with a simple linear mapping and use this approach for cross-lingual transfer. We also observe that the model obtains the highest transfer performance with the middle layer representation alignment, and not the last layers. The performance gap between monolingual MLM alignment and bilingual MLM is higher in NER compared to parsing, suggesting the syntactic information needed for parsing might be easier to align with a simple mapping while entity information requires more explicit entity alignment.

5.1.3 Sentence-level alignment

In this case, X and Y are obtained by average pooling subword representation (excluding special token) of sentences *at each layer* of monolingual BERT. We use multi-way parallel sentences from XNLI for alignment supervision and Tatoeba (Schwenk et al., 2019) for evaluation.

[2] In Appendix A, we also present the same alignment step with biMLM but only observed improvement in parsing.



(a) Non-contextual word embeddings alignment

(b) Contextual word embedding alignment

Figure 4: Alignment of word-level representations from monolingual BERT models on subset of MUSE benchmark. Figure 4a and Figure 4b are not comparable due to different embedding vocabularies.

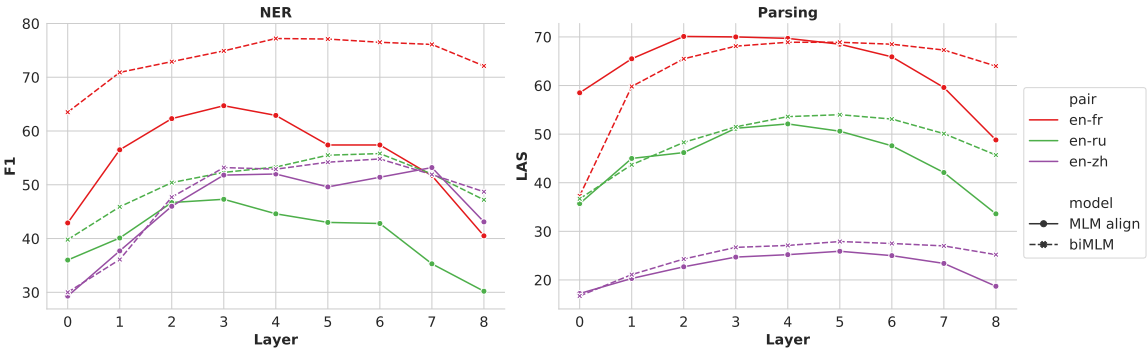


Figure 5: Contextual representation alignment of different layers for zero-shot cross-lingual transfer.

Figure 6 shows the sentence similarity search results with nearest neighbor search and cosine similarity, evaluated by precision at 1, with four language pairs. Here the best result is obtained at lower layers. The performance is surprisingly good given we only use 10k parallel sentences to learn the alignment without fine-tuning at all. As a reference, the state-of-the-art performance is over 95%, obtained by LASER (Artetxe and Schwenk, 2019) trained with millions of parallel sentences.

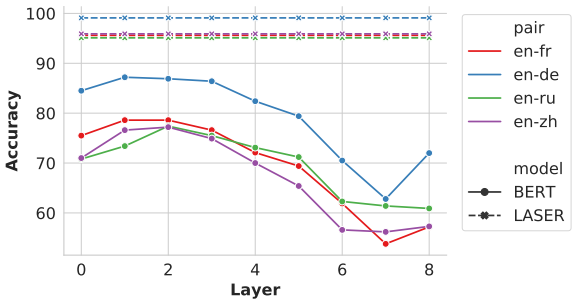


Figure 6: Parallel sentence retrieval accuracy after Procrustes alignment of monolingual BERT models.

5.1.4 Conclusion

These findings demonstrate that both word-level, contextual word-level, and sentence-level BERT representations can be aligned with a simple orthogonal mapping. Similar to the alignment of word embeddings (Mikolov et al., 2013), this shows that BERT models are similar across languages. This result gives more intuition on why mere parameter sharing is sufficient for multilingual representations to emerge in multilingual masked language models.

5.2 Neural network similarity

Based on the work of Kornblith et al. (2019), we examine the centered kernel alignment (CKA), a neural network similarity index that improves upon canonical correlation analysis (CCA), and use it to measure the similarity across both monolingual and bilingual masked language models. The linear CKA is both invariant to orthogonal transformation and isotropic scaling, but are not invertible to any linear transform. The linear CKA similarity measure is defined as follows:

$$\mathrm{CKA}(X, Y) = \frac{\|Y^T X\|_{\mathrm{F}}^2}{(\|X^T X\|_{\mathrm{F}} \|Y^T Y\|_{\mathrm{F}})},$$

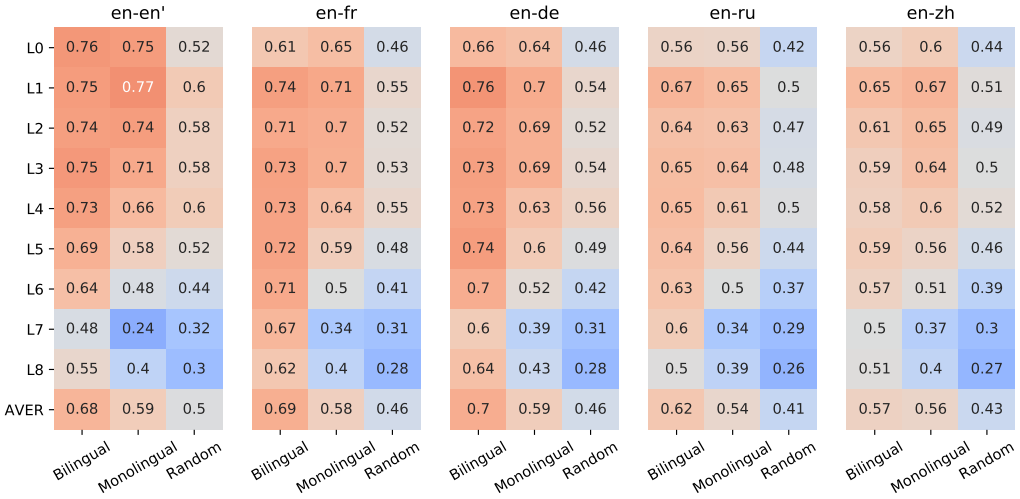


Figure 7: CKA similarity of mean-pooled multi-way parallel sentence representation at each layers. Note en′ corresponds to paraphrases of en obtained from back-translation (en-fr-en′). Random encoder is only used by non-Englih sentences. L0 is the embeddings layers while L1 to L8 are the corresponding transformer layers. The average row is the average of 9 (L0-L8) similarity measurements.

where X and Y correspond respectively to the matrix of the d-dimensional mean-pooled (excluding special token) subword representations at layer l of the n parallel source and target sentences.

In Figure 7, we show the CKA similarity of monolingual models, compared with bilingual models and random encoders, of multi-way parallel sentences (Conneau et al., 2018) for five languages pair: English to English′ (obtained by back-translation from French), French, German, Russian, and Chinese. The monolingual en′ is trained on the same data as en but with different random seed and the bilingual en-en′ is trained on English data but with separate embeddings matrix as in §4.3. The rest of the bilingual MLM is trained with the Default setting. We only use random encoder for non-English sentences.

Figure 7 shows bilingual models have slightly higher similarity compared to monolingual models with random encoders serving as a lower bound. Despite the slightly lower similarity between monolingual models, it still explains the alignment performance in §5.1. Because the measurement is also invariant to orthogonal mapping, the CKA similarity is highly correlated with the sentence-level alignment performance in Figure 6 with over 0.9 Pearson correlation for all four languages pairs. For monolingual and bilingual models, the first few layers have the highest similarity, which explains why Wu and Dredze (2019) finds freezing bottom layers of mBERT helps cross-lingual transfer. The similarity gap between monolingual model and bilingual model decrease as the languages pair become more distant. In other words, when languages are similar, using the same model increase representation similarity. On the other hand, when languages are dissimilar, using the same model does not help representation similarity much. Future work could consider how to best train multilingual models covering distantly related languages.

6 Discussion

In this paper, we show that multilingual representations can emerge from unsupervised multilingual masked language models with only parameter sharing of some Transformer layers. Even without any anchor points, the model can still learn to map representations coming from different languages in a single shared embedding space. We also show that isomorphic embedding spaces emerge from monolingual masked language models in different languages, similar to word2vec embedding spaces (Mikolov et al., 2013). By using a linear mapping, we are able to align the embedding layers and the contextual representations of Transformers trained in different languages. We also use the CKA neural network similarity index to probe the similarity between BERT Models and show that the early layers of the Transformers are more similar across languages than the last layers. All of these effects were stronger for more closely related languages, suggesting there is room for significant improvements on more distant language pairs.