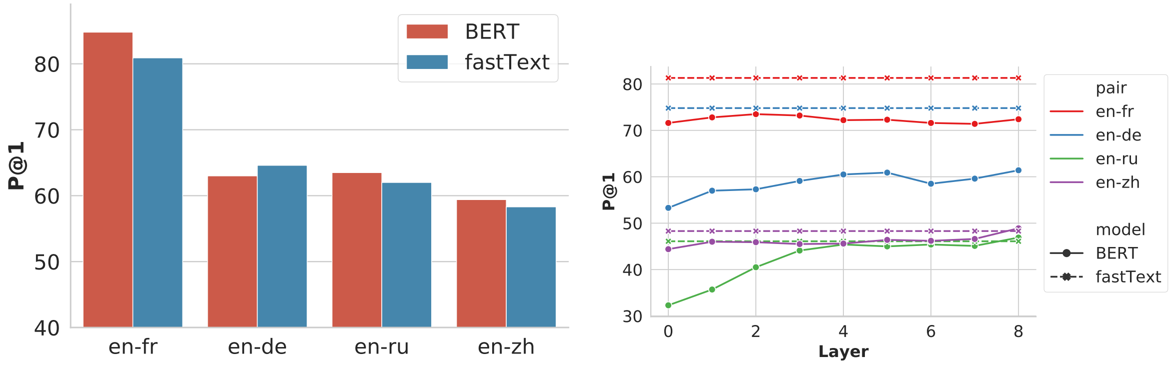

(a) Non-contextual word embeddings alignment

(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.

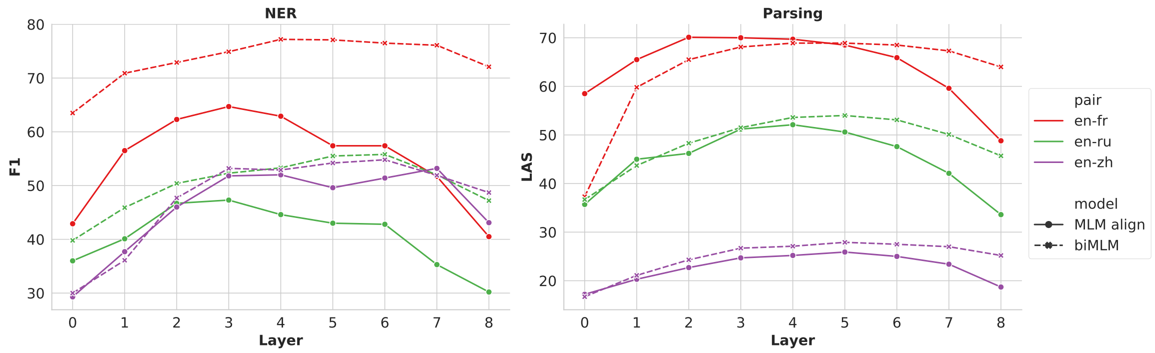

Figure 5: Contextual representation alignment of different layers for zero-shot cross-lingual transfer.

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.

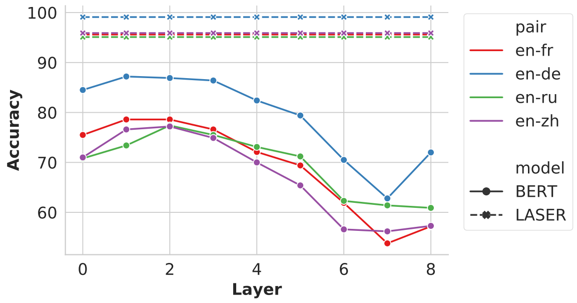

Figure 6: Parallel sentence retrieval accuracy after Procrustes alignment of monolingual BERT models.

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:

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