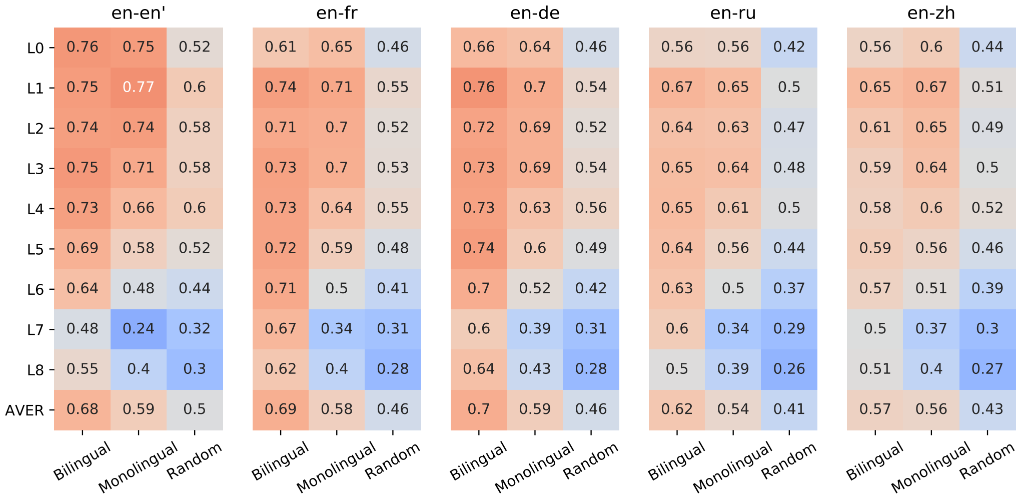

	en-en'			en-fr			en-de			en-ru			en-zh		
	Bilingual	Monolingual	Random	Bilingual	Monolingual	Random	Bilingual	Monolingual	Random	Bilingual	Monolingual	Random	Bilingual	Monolingual	Random
L0	0.76	0.75	0.52	0.61	0.65	0.46	0.66	0.64	0.46	0.56	0.56	0.42	0.56	0.6	0.44
L1	0.75	0.77	0.6	0.74	0.71	0.55	0.76	0.7	0.54	0.67	0.65	0.5	0.65	0.67	0.51
L2	0.74	0.74	0.58	0.71	0.7	0.52	0.72	0.69	0.52	0.64	0.63	0.47	0.61	0.65	0.49
L3	0.75	0.71	0.58	0.73	0.7	0.53	0.73	0.69	0.54	0.65	0.64	0.48	0.59	0.64	0.5
L4	0.73	0.66	0.6	0.73	0.64	0.55	0.73	0.63	0.56	0.65	0.61	0.5	0.58	0.6	0.52
L5	0.69	0.58	0.52	0.72	0.59	0.48	0.74	0.6	0.49	0.64	0.56	0.44	0.59	0.56	0.46
L6	0.64	0.48	0.44	0.71	0.5	0.41	0.7	0.52	0.42	0.63	0.5	0.37	0.57	0.51	0.39
L7	0.48	0.24	0.32	0.67	0.34	0.31	0.6	0.39	0.31	0.6	0.34	0.29	0.5	0.37	0.3
L8	0.55	0.4	0.3	0.62	0.4	0.28	0.64	0.43	0.28	0.5	0.39	0.26	0.51	0.4	0.27
AVER	0.68	0.59	0.5	0.69	0.58	0.46	0.7	0.59	0.46	0.62	0.54	0.41	0.57	0.56	0.43

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-Engligh 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.