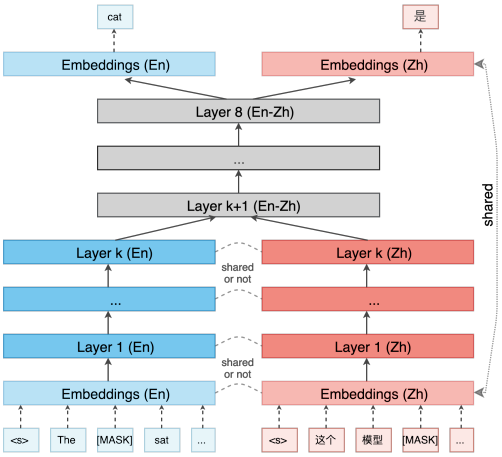

Figure 1: **On the impact of anchor points and parameter sharing on the emergence of multilingual representations.** We train bilingual masked language models and remove parameter sharing for the embedding layers and first few Transformers layers to probe the impact of anchor points and shared structure on cross-lingual transfer.

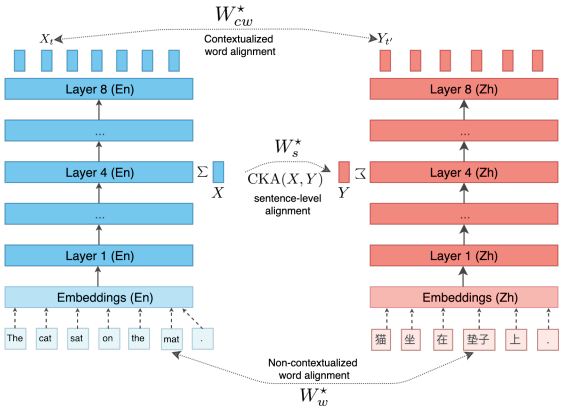

Figure 2: **Probing the layer similarity of monolingual BERT models.** We investigate the similarity of separate monolingual BERT models at different levels. We use an orthogonal mapping between the pooled representations of each model. We also quantify the similarity using the centered kernel alignment (CKA) similarity index.

classifier, which takes the representation of the first subword of each word as input. We report span-level F1. We adopt a simple post-processing heuristic to obtain a valid span, rewriting standalone `I-X` into `B-X` and `B-X I-Y I-Z` into `B-Z I-Z I-Z`, following the final entity type. We report the span-level F1.

Parsing Finally, we use the Universal Dependencies (UD v2.3) (Nivre, 2018) for dependency parsing. We consider the following four treebanks: English-EWT, French-GSD, Russian-GSD, and Chinese-GSD. The task-specific layer is a graph-based parser (Dozat and Manning, 2016), using representations of the first subword of each word as inputs. We measure performance with the labeled attachment score (LAS).

4 Dissecting mBERT/XLM models

We hypothesize that the following factors play important roles in what makes multilingual BERT multilingual: domain similarity, shared vocabulary (or anchor points), shared parameters, and language similarity. Without loss of generality, we focus on bilingual MLM. We consider three pairs of languages: English-French, English-Russian, and English-Chinese.

4.1 Domain Similarity

Multilingual BERT and XLM are trained on the Wikipedia comparable corpora. Domain similarity has been shown to affect the quality of cross-lingual word embeddings (Conneau et al., 2017), but this effect is not well established for masked language models. We consider domain difference by training on Wikipedia for English and a random subset of Common Crawl of the same size for the other languages (**Wiki-CC**). We also consider a model trained with Wikipedia only (**Default**) for comparison.

The first group in Tab. 1 shows domain mismatch has a relatively modest effect on performance. XNLI and parsing performance drop around 2 points while NER drops over 6 points for all languages on average. One possible reason is that the labeled WikiAnn data for NER consists of Wikipedia text; domain differences between source and target language during pretraining hurt performance more. Indeed for English and Chinese NER, where neither side comes from Wikipedia, performance only drops around 2 points.

4.2 Anchor points

Anchor points are *identical strings* that appear in both languages in the training corpus. Translingual words like *DNA* or *Paris* appear in the Wikipedia of many languages with the same meaning. In mBERT, anchor points are naturally preserved due to joint BPE and shared vocabulary across languages. Anchor point existence has been suggested as a key ingredient for effective cross-lingual transfer since they allow the shared encoder to have at least some direct tying of meaning across different languages (Lample and Conneau, 2019; Pires et al., 2019; Wu and Dredze, 2019). However, this effect

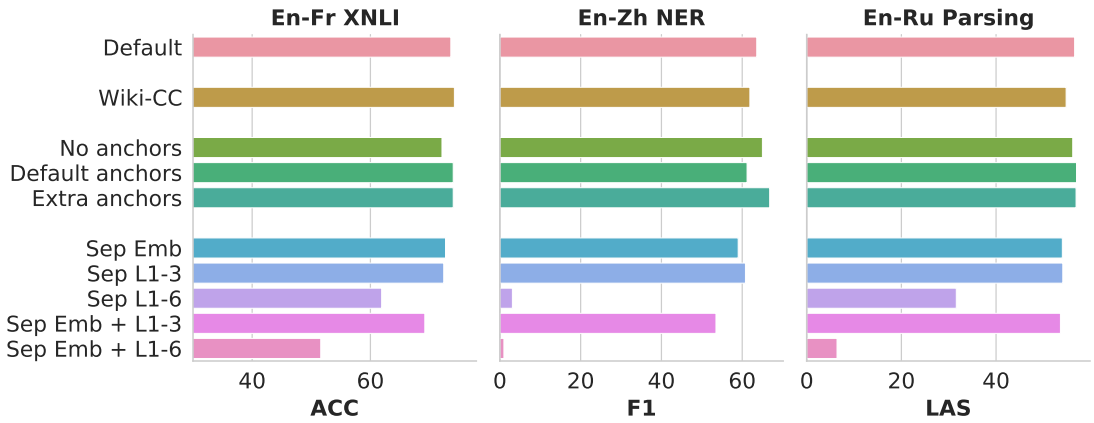

Figure 3: Cross-lingual transfer of bilingual MLM on three tasks and language pairs under different settings. Others tasks and languages pairs follows similar trend. See Tab. 1 for full results.

Model	Domain	BPE Merges	Anchors Pts	Share Param.	Softmax	XNLI (Acc) fr	ru	zh	Δ	NER (F1) fr	ru	zh	Δ	Parsing (LAS) fr	ru	zh	Δ
Default	Wiki-Wiki	80k	all	all	shared	73.6	68.7	68.3	0.0	79.8	60.9	63.6	0.0	73.2	56.6	28.8	0.0
Domain Similarity (§4.1)																	
Wiki-CC	Wiki-CC	-	-	-	-	74.2	65.8	66.5	-1.4	74.0	49.6	61.9	-6.2	71.3	54.8	25.2	-2.5
Anchor Points (§4.2)																	
No anchors	-	40k/40k	0	-	-	72.1	67.5	67.7	-1.1	74.0	57.9	65.0	-2.4	72.3	56.2	27.4	-0.9
Default anchors	-	40k/40k	-	-	-	74.0	68.1	68.9	+0.1	76.8	56.3	61.2	-3.3	73.0	57.0	28.3	-0.1
Extra anchors	-	-	extra	-	-	74.0	69.8	72.1	+1.8	76.1	59.7	66.8	-0.5	73.3	56.9	29.2	+0.3
Parameter Sharing (§4.3)																	
Sep Emb	-	40k/40k	0*	Sep Emb	lang-specific	72.7	63.6	60.8	-4.5	75.5	57.5	59.0	-4.1	71.7	54.0	27.5	-1.8
Sep L1-3	-	40k/40k	-	Sep L1-3	-	72.4	65.0	63.1	-3.4	74.0	53.3	60.8	-5.3	69.7	54.1	26.4	-2.8
Sep L1-6	-	40k/40k	-	Sep L1-6	-	61.9	43.6	37.4	-22.6	61.2	23.7	3.1	-38.7	61.7	31.6	12.0	-17.8
Sep Emb + L1-3	-	40k/40k	0*	Sep Emb + L1-3	lang-specific	69.2	61.7	56.4	-7.8	73.8	46.8	53.5	-10.0	68.2	53.6	23.9	-4.3
Sep Emb + L1-6	-	40k/40k	0*	Sep Emb + L1-6	lang-specific	51.6	35.8	34.4	-29.6	56.5	5.4	1.0	-47.1	50.9	6.4	1.5	-33.3

Table 1: Dissecting bilingual MLM based on zero-shot cross-lingual transfer performance. - denote the same as the first row (**Default**). Δ denote the difference of average task performance between a model and **Default**.

has not been carefully measured.

We present a controlled study of the impact of anchor points on cross-lingual transfer performance by varying the amount of shared subword vocabulary across languages. Instead of using a single joint BPE with 80k merges, we use language-specific BPE with 40k merges for each language. We then build vocabulary by taking the union of the vocabulary of two languages and train a bilingual MLM (**Default anchors**). To remove anchor points, we add a language prefix to each word in the vocabulary before taking the union. Bilingual MLM (**No anchors**) trained with such data has no shared vocabulary across languages. However, it still has a single softmax prediction layer shared across languages and tied with input embeddings.

As Wu and Dredze (2019) suggest there may also be correlation between cross-lingual performance and anchor points, we additionally increase anchor points by using a bilingual dictionary to create code switch data for training bilingual MLM (**Extra anchors**). For two languages, ℓ_1 and ℓ_2, with bilingual dictionary entries d_{ℓ_1,ℓ_2}, we add anchors to the training data as follows. For each training word w_{ℓ_1} in the bilingual dictionary, we either leave it as is (70% of the time) or randomly replace it with one of the possible translations from the dictionary (30% of the time). We change at most 15% of the words in a batch and sample word translations from PanLex (Kamholz et al., 2014) bilingual dictionaries, weighted according to their translation quality [1].

The second group of Tab. 1 shows cross-lingual transfer performance under the three anchor point conditions. Anchor points have a clear effect on performance and more anchor points help, especially in the less closely related language pairs (e.g. English-Chinese has a larger effect than English-French with over 3 points improvement on NER and XNLI). However, surprisingly, effective transfer is still possible with no anchor points. Com-

[1] Although we only consider pairs of languages, this procedure naturally scales to multiple languages, which could produce larger gains in future work.

paring no anchors and default anchors, the performance of XNLI and parsing drops only around 1 point while NER even improve 1 points averaging over three languages. Overall, these results show that we have previously overestimated the contribution of anchor points during multilingual pretraining. Concurrently, Karthikeyan et al. (2020) similarly find anchor points play minor role in learning cross-lingual representation.

4.3 Parameter sharing

Given that anchor points are not required for transfer, a natural next question is the extent to which we need to tie the parameters of the transformer layers. Sharing the parameters of the top layer is necessary to provide shared inputs to the task-specific layer. However, as seen in Figure 1, we can progressively separate the *bottom* layers 1:3 and 1:6 of the Transformers and/or the embedding layers (including positional embeddings) (**Sep Emb**; **Sep L1-3**; **Sep L1-6**; **Sep Emb + L1-3**; **Sep Emb + L1-6**). Since the prediction layer is tied with the embeddings layer, separating the embeddings layer also introduces a language-specific softmax prediction layer for the cloze task. Additionally, we only sample random words within one language during the MLM pretraining. During fine-tuning on the English training set, we freeze the language-specific layers and only fine-tune the shared layers.

The third group in Tab. 1 shows cross-lingual transfer performance under different parameter sharing conditions with "Sep" denote which layers **is not** shared across languages. Sep Emb (effectively no anchor point) drops more than No anchors with 3 points on XNLI and around 1 point on NER and parsing, suggesting have a cross-language softmax layer also helps to learn cross-lingual representations. Performance degrades as fewer layers are shared for all pairs, and again the less closely related language pairs lose the most. Most notably, the cross-lingual transfer performance drops to random when separating embeddings and bottom 6 layers of the transformer. However, reasonably strong levels of transfer are still possible without tying the bottom three layers. These trends suggest that parameter sharing is the key ingredient that enables the learning of an effective cross-lingual representation space, and having language-specific capacity does not help learn a language-specific encoder for cross-lingual representation. Our hypothesis is that the representations that the models learn for different languages are similarly shaped and models can reduce their capacity budget by aligning representations for text that has similar meaning across languages.

4.4 Language Similarity

Finally, in contrast to many of the experiments above, language similarity seems to be quite important for effective transfer. Looking at Tab. 1 column by column in each task, we observe performance drops as language pairs become more distantly related. Using extra anchor points helps to close the gap. However, the more complex tasks seem to have larger performance gaps and having language-specific capacity does not seem to be the solution. Future work could consider scaling the model with more data and cross-lingual signal to close the performance gap.

4.5 Conclusion

Summarised by Figure 3, parameter sharing is the most important factor. More anchor points help but anchor points and shared softmax projection parameters are not necessary for effective cross-lingual transfer. Joint BPE and domain similarity contribute a little in learning cross-lingual representation.

5 Similarity of BERT Models

To better understand the robust transfer effects of the last section, we show that independently trained monolingual BERT models learn representations that are similar across languages, much like the widely observed similarities in word embedding spaces. In this section, we show that independent monolingual BERT models produce highly similar representations when evaluated at the word level (§5.1.1), contextual word-level (§5.1.2), and sentence level (§5.1.3) . We also plot the cross-lingual similarity of neural network activation with center kernel alignment (§5.2) at each layer. We consider five languages: English, French, German, Russian, and Chinese.

5.1 Aligning Monolingual BERTs

To measure similarity, we learn an orthogonal mapping using the Procrustes (Smith et al., 2017) approach:

$$W^\star = \underset{W \in O_d(\mathbb{R})}{\operatorname{argmin}} \|WX - Y\|_F = UV^T$$