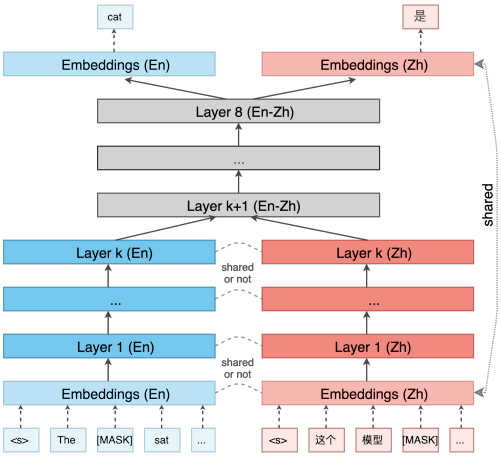

Figure 1: **On the impact of anchor points and parameter sharing on the emergence of multilingual representations.** We train bilingual masked language models and remove parameter sharing for the embedding layers and first few Transformers layers to probe the impact of anchor points and shared structure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

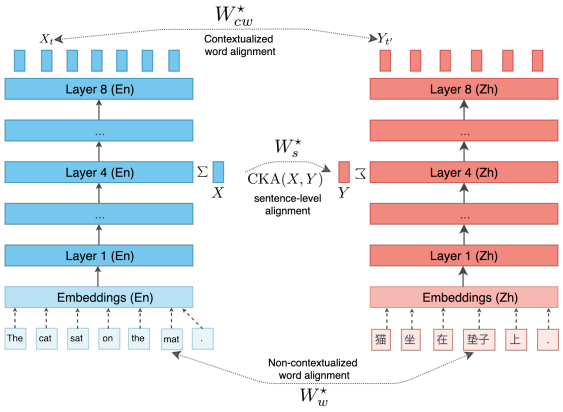

Figure 2: **Probing the layer similarity of monolingual BERT models.** We investigate the similarity of separate monolingual BERT models at different levels. We use an orthogonal mapping between the pooled representations of each model. We also quantify the similarity using the centered kernel alignment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