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.